Information and Communication Technologies

EPIWORK

Developing the Framework for an Epidemic Forecast Infrastructure

http://www.epiwork.eu

Project no. 231807

D3.6 - Report: Final Specification of the Epidemic Marketplace Platform and Evaluation Results

---

Period covered:
Start date of project: February 1st, 2009
Due date of deliverable: July 31st 2013
Distribution: Public

Date of preparation: July 31st, 2013
Duration:
Actual submission date: July 31th, 2013
Status:

Project Coordinator: Alessandro Vespignani
Project Coordinator Organisation Name: ISI Foundation
Lead contractor for this deliverable: FFCUL

Work package participants

The following partners have taken active part in the work leading to the elaboration of this document, even if they might not have directly contributed writing parts of this document:

João Zamite, João D. Ferreira, Paulo Graça, Carlos Santos, Tiago Posse. Cátia Pesquita, Dulce Domingos, Francisco Couto and Mário J. Silva.

Change log

Version	Date	Amended by	Changes
0.1	2013-01-20	Mário J. Silva	First draft
1.0	2013-07-31	Mário J. Silva	Publication

D3.6 - Report: Final Specification of the Epidemic Marketplace Platform and Evaluation Results

João Zamite[1], João D. Ferreira[1], Paulo Graça[1], Carlos Santos[1], Tiago Posse[1], Cátia Pesquita[1], Dulce Domingos[1], Francisco Couto[1], and Mário J. Silva[2]

[1]University of Lisbon, Faculty of Sciences, LASIGE, Portugal

[2]University of Lisbon, IST/INESC-ID, LASIGE, Portugal

July, 2013

Abstract

This document describes the final version of the Epidemic Marketplace (EM), as deployed at `http://epimarketplace.net`. Developed under the EPIWORK project, the EM incorporates research and development work on a model for representing epidemic datasets and its implementation as the platform's software and collaborative computing infrastructure. This includes the final user interface design, newly developed in the last reporting period. The report gives an account of the research work on the use of ontologies for annotation of epidemiological resources and its embodying within the EM metadata model. It also includes a description of the approaches used to bring new resources to the Epidemic Marketplace, namely through the integration of data from an epidemics simulation platform, monitorization data and also data collected by the user community. Finally, the report also discusses strategies for recovering from potential issues with the existing platform and presents the implemented monitoring and alert systems for fault handling. A summary of platform usage statistics, which can also be monitored interactively at the platform, is also provided.

Contents

Chapter 1

Introduction

The recent explosion in produced data in every field of science has resulted in the emergence of highly data driven sciences. These sciences, in turn, are making extensive use of computational simulations, which also generate new data resources, adding to the increasing complexity of storing, managing and sharing needs [1].

For this *data tsunami* to be useful, it is necessary that the generated data is adequately managed, so it can be re-used to create more knowledge or influence decision making. To accomplish this, adequate information platforms that can deal with the increasing volume, scale and complexity required to managing, preserving and sharing these data are necessary.

In 2010, the Riding the Wave report identified the need for an e-infrastructure for data management enabling seamless access, re-use and trust of data [2]. The authors present challenges to improve scientific discovery and support collaboration across disciplinary and geographical boundaries. The challenges include:

- data preservation and curation;

- linking people and data;
- describing data through adequate metadata and semantics for data discovery;
- enabling interoperability and data exchange across scientific domains; and
- establishing trust through adequate authentication and authorization platforms.

These challenges are directly linked with the role of information platforms for data driven sciences. For resources to be used in scientific discovery in a collaborative environment it is necessary to ensure that data is preserved, discoverable, accurately described, shareable and also secure. These are the requirements an information system must meet in order to make data available and foster viable data driven research.

Epidemiology is also, by nature, a data-intensive research field, combining results from empirical, analytical and simulation studies based on large volumes of data in which timeliness and accuracy in the analyses are crucial. Furthermore, the recent explosion of mobile phone and Internet usage is creating an immense trove of data, containing epidemiologically relevant behaviors, such as deciding on preventive measures and treatment choices, as well as reporting disease symptoms [3]. Finally, data driven models are now generating libraries of scenarios and patterns that need to be fed and contrasted with real world datasets [4] [5].

Scientific articles in epidemiology usually do not provide direct references to where and how the datasets supporting their findings can be obtained. In contrast, this is a standard procedure in molecular biology, where authors are strongly encouraged to deposit their datasets prior to article publication

in public databases, such as GenBank and UniProt. This grants readers a deeper insight into the conducted research and promotes repeatability of experiments and reuse of data. However, in epidemiology the complexity of finding datasets makes the direct replication of epidemiological studies non-trivial or in many cases even impossible, which is of course a serious bottleneck to the scientific progress in this area.

Epidemiologic studies require not only raw mobility, surveillance, demographic, economic, social network data, but also different kinds of infomation resources, such as models and simulations. In addition, Epidemiology is a multi-disciplinary discipline integrating diverse areas of knowledge, such as medicine, biology, statistics, social sciences and geography. Thus, epidemiological studies would much benefit from the new scientific methodology designated as the fourth paradigm, which addresses the challenges raised from our need to validate, analyze, visualize, store, and curate the large amounts of generated data [1]. An information platform for epidemiology must thus be able to manage all these types of data while aiming at promoting collaboration and interoperability.

This report presents the final results of the development of an epidemiological information platform within EPIWORK, comprising the research and implementation done during this period and presents the final specification of the resulting platform, the Epidemic Marketplace (EM), which addresses the above challenges in the epidemiology domain [6]. The EM stores and manages health-related resources, from research papers and surveillance reports to sensitive data, such as epidemic incidence datasets. It has been publicly available to the research community since September 2010.

The EM is fully built on open source technologies. The back end is based on a Fedora Commons repository [7] for storing and managing resources

and a Lightweight Directory Access Protocol (LDAP) server [8] for user management. The EM provides a set of web services that enable full access and manipulation of the repository content, and a Drupal-based user interface front-end [9] for interactive upload and manipulation of resources. The front-end uses the same web services that are offered to external applications. The whole system has been running on a cluster of Linux machines.

The EM aims at encouraging epidemiological data sharing to facilitate data-intensive research. Collaboration among EPIWORK partners has shown the need to standardize metadata information of digital resources. For this reason, the EM has established a metadata model for the annotation of epidemic resources, which is an extended a generic metadata model based on the interoperable Dublin Core (DC) metadata standard [10]. The EM metadata model provides a general semantic vocabulary for characterizing the metadata of online resources. The DC standard only provides a structure for organising the meta-data and only models generic aspects, such as time and location or access rights. In EPIWORK, we researched how to effectively perform the annotation of epidemiological resources with a semantic network that would provide, for each metadata field, valid names which could be used. The approach taken by EPIWORK involved creating a Network of Epidemiology-Related Ontologies (NERO) [11], which includes concepts which are relevant of epidemiological resource categorization, such as diseases and vaccines. The benefit of using ontology-based controlled vocabularies is that the EM takes the data closer to the vision a Web of Knowledge as opposed to a Web of Text [12]. The expressiveness of these ontologies enable users to accurately express the contents of the resources while enabling Semantic Web tools to be used on these resources, supporting powerful queries or a semantic analysis that can be used to draw inferences based on

these annotations [13]. Additionally, such tools can be used to facilitate the creation of metadata by suggesting the user a set of concepts for annotation based on the contents of a resource being uploaded to the EM.

An additional challenge when coping with epidemiological data is that the data itself is often sensitive in nature and must therefore be shared under adequate access control mechanisms. While some previous approaches hint towards giving resource owners more control over their resources, when managing sensitive health-related resources, the standard practice is to give this responsibility to the resource owner, which is, in some countries, a legal requirement. The development of the EM also involved identifying the access control requirements of a repository for epidemic resources and, based on the elicited requirements, proposing a group-based approach to access control over repository resources [14]. We adopted a decentralized and discretionary approach over permission assignment as well as in the management of user groups. Additionally, our access control model includes an object structure that separates data from meta-data, enabling the search of resources without exposing sensitive data. We evaluate the applicability of the proposed model on the EM, built from open source technologies showing it is a feasible solution for similar scientific data management environments.

The EM is built on open source technologies. The back end is based on a Fedora Commons repository [7] for storing and managing resources and a Lightweight Directory Access Protocol (LDAP) server [8] for user management. The EM also includes a Drupal-based front-end [9] and web services to access repository content, both by the front-end and client applications.

We developed the EM from the start for being used both directly by humans and client applications. Because epidemiologists often use simulation

software as well as other automated software to perform data analysis, it is necessary that the EM provides an application programming interface (API) that can be used by that software to pull data sets directly. This fosters integration and re-use of epidemiological data by enabling applications to retrieve fresh data from the EM. It also fosters collaboration by providing a plaftorm for sharing resources that enables both automatic uploading and easy user uploading of new data and other research results to the platform.

Changing the habits of the epidemic modelers community to adopting data sharing practices through a platform like the Epidemic Marketplace will necessarily take time and persistence. To observe the impact of the information platform we need tools to effectively monitor the usage of the platform. The EM also includes a module for collecting usage statistics, in terms of resources, registered users and user accesses over time.

The remainder of this report describes the various components of the EM in detail, their development and the obtained results. It is organised as follows: Chapter 2 presents the system architecture of the EM; Chapter 3 describes the hardware infrastructure and how services are deployed on this architecture, as well as the challenges and solutions to distribute and maximize the platform availability; Chapter 4 presents the development of the latest stage of the user interface and the insight behind it; Chapter 5 presents the effort towards the semantification of the stored epidemiological resources; Chapter 6 describes data harvesting strategies and partnerships with other work packages to populate and promote the use of the data repository. Chapter7 presents the results of the development of the EM and usage statistics; Chapter 8 presents the conclusions of this report.

1.1 Publications

The development of the EM spanned a period of four and a half years, during which multiple publications and presentations have presented most of the work reported here with much more detail. We list below all the publications documenting the EM that we have produced:

[1] J. Zamite, D. Domingos, M. J. Silva, and C. Santos, "Group-based discretionary access control for epidemiological resources," in *HCist 2013 - International Conference on Health and Social Care Information Systems and Technologies*, ser. Procedia Technology. Elsevier, October 2013.

[2] C. Santos, D. Domingos, J. Zamite, P. Graça, and M. J. Silva, "Access control for shared epidemic datasets," Poster @ Epiwork International Workshop "Digital Epidemiology", Turin, Italy, 2013.

[3] C. Pesquita, J. D. Ferreira, F. M. Couto, and M. J. Silva, "Semi-automated annotation of epidemiological resources," Poster @ Epiwork International Workshop "Digital Epidemiology", Turin, Italy, 2013.

[4] T. Grego, F. Pinto, and F. Couto, "Lasige: using conditional random fields and chebi ontology," in *Proceedings of the International Workshop on Semantic Evaluation (SemEval2013)*, 2013. [Online]. Available: http://www.cs.york.ac.uk/semeval-2013/accepted/71_Paper.pdf

[5] T. Grego and F. Couto, "Enhancement of chemical entity identification in text using semantic similarity validation," *PLOS ONE*, vol. 8, no. 5, p. e62984, 2013. [Online]. Available: http://dx.plos.org/10.1371/journal.pone.0062984

[6] J. Ferreira, D. Paolotti, F. Couto, and M. J. Silva, "On the usefulness of ontologies in epidemiologic research and practice," *Journal of Epidemiology and Community Health*, 2012. [Online]. Available: http://dx.doi.org/10.1136/jech-2012-201142

[7] C. Pesquita and F. M. Couto, "Predicting the extension of biomedical ontologies," *PLOS Computational Biology*, vol. 8, no. 9, p. e1002630, 2012. [Online]. Available: dx.doi.org/10.1371/journal.pcbi.1002630

[8] J. D. Ferreira and F. M. Couto, "Semantic similarity in the biomedical domain," Poster @ Bioinformatics Open Days 2012, Braga, Portugal, 2012.

[9] F. M. Couto, J. D. Ferreira, J. Zamite, C. Santos, T. Posse, P. Graça, D. Domingos, and M. J. Silva, "The epidemic marketplace platform: towards semantic characterization of epidemiological resources using biomedical ontologies," in *International Conference on Biomedical Ontologies (ICBO)*, 2012.

[10] J. D. Ferreira, C. Pesquita, F. M. Couto, and M. J. Silva, "Bringing epidemiology into the semantic web," in *International Conference on Biomedical Ontologies (ICBO)*, 2012.

[11] C. Gioannini and J. Zamite, "Integrating the gleamviz simulator tool with the epidemic marketplace platform," Poster presented at EE2, Epiwork/Epifor 2nd International Workshop: Facing the Challenge of Infectious Diseases, 2012. [Online]. Available: http://www.isi.it/events/ee2/

[12] J. Zamite, D. Domingos, and M. J. Silva, "Owner-centred group-based access control for epidemic resources," Poster presented

at EE2, Epiwork/Epifor 2nd International Workshop: Facing the Challenge of Infectious Diseases, 2012. [Online]. Available: http://www.isi.it/events/ee2/

[13] J. D. Ferreira, F. M. Couto, and M. J. Silva, "Ontologies in the epidemiological domain," Poster presented at EE2, Epiwork/Epifor 2nd International Workshop: Facing the Challenge of Infectious Diseases, 2012. [Online]. Available: http://www.isi.it/events/ee2/

[14] J. D. Ferreira, C. Pesquita, F. Couto, and M. J. Silva, "Epiwork deliverable 3.5: Epidemic data ontology," University of Lisbon, Faculty of Sciences, LASIGE, Tech. Rep., January 2012.

[15] C. P. F. Sousa, "Epidemic marketplace: Repositório e web services," Master's thesis, University of Lisbon, Faculty of Sciences, January 2012.

[16] J. Zamite, F. Silva, F. Couto, and M. Silva, "Medcollector: Multisource epidemic data collector," *Transactions on Large-scale Data-and Knowledge-centered Systems IV: Special Issue on Database Systems for Biomedical Applications*, vol. 6990, pp. 40–72, 2011.

[17] F. Couto and M. J. Silva, "Disjunctive shared information between ontology concepts: application to gene ontology," *Journal of Biomedical Semantics*, vol. 2, no. 5, 2011. [Online]. Available: http://www.jbiomedsem.com/content/2/1/5/abstract

[18] J. Duque, "Mediação dados-informação: Design de informação para a epidemic marketplace," Master's thesis, University of Lisbon, School of Fine Arts, November 2011.

[19] C. Pesquita and F. Couto, "Where go is going and what it means for ontology extension," in *International Conference on Biomedical Ontologies*, 2011.

[20] B. Tavares, H. Bastos, D. Faria, J. D. Ferreira, T. Grego, C. Pesquita, and F. Couto, "The biomedical ontology applications (boa) framework," in *Proceedings of ICBO*, 2011.

[21] J. D. Ferreira and F. Couto, "Generic semantic relatedness measure for biomedical ontologies," in *Proceedings of ICBO*, 2011.

[22] H. Ferreira, "O mediador do epidemic marketplace," Master's thesis, University of Lisbon, Faculty of Sciences, February 2011.

[23] M. J. Silva, F. Couto, D. Domingos, J. Duque, H. Ferreira, L. F. Lopes, D. Paolotti, F. Silva, P. Sousa, and J. Zamite, "D 3.3 public release of the epidemic marketplace platform," LASIGE, University of Lisbon, Faculty of Sciences, Tech. Rep., September 2010.

[24] L. F. Lopes, "A metadata model for the annotation of epidemiological data," Master's thesis, University of Lisbon, Faculty of Sciences, September 2010.

[25] J. Zamite, "Multisource epidemic data collector," Master's thesis, University of Lisbon, Faculty of Sciences, September 2010.

[26] F. A. Silva, M. J. Silva, and F. Couto, "Epidemic marketplace: an e-science platform for epidemic modelling and analysis," *ERCIM News*, no. 82, pp. 43–44, July 2010, special Theme: Computational Biology.

[27] L. F. Lopes, F. A. Silva, F. Couto, J. Zamite, H. Ferreira, C. Sousa, and M. J. Silva, "Epidemic marketplace: An information management

system for epidemiological data." in *Proceedings of the ITBAM - DEXA 2010*, 2010.

[28] J. Zamite, F. A. Silva, F. Couto, and M. J. Silva, "Medcollector: Multisource epidemic data collector," in *Proceedings of ITBAM'10 - 1st International Conference on Information Technology in Bio- and Medical Informatics - DEXA 2010*, August 2010.

[29] M. J. Silva, F. A. Silva, L. F. Lopes, and F. Couto, "Building a digital library for epidemic modelling," in *Proceedings of ICDL 2010 - The International Conference on Digital Libraries*, vol. 1. New Delhi, India: TERI Press – New Delhi, India, 23–27 February 2010, invited Paper.

[30] L. F. Lopes, J. Zamite, B. Tavares, F. Couto, F. A. Silva, and M. J. Silva, "Automated social network epidemic data collector," in *INForum - Simpósio de Informática*, September 2009.

Chapter 2

System Architecture

The EM was envisioned as an information platform capable of storing and managing data produced by other work packages as well as epidemiologists external to EPIWORK. We developed an information platform to mediate access to distributed collections of public health data, offering an easy and safe way to share data for those data providers who want to collaborate with epidemiological modelers.

Researchers use this platform in multiple ways, but the appearance is that of a catalogue of data sources containing the metadata describing existing databases, to publish information about their own data or as the host of mediating software that can automatically process queries for epidemiological data available from the information sources connected to the platform.

This chapter starts with a presentation of the requirements for an epidemiological platform. We will then explore the software architecture solution that has been developed.

2.1 Requirements

Epidemiology is data-driven and uses multiple types of data from various sources. A single epidemiological study may use heterogeneous data from demography data (such as age distribution), environmental data, social data, mobility data (such as airport traffic), genetic data, among others, each of these with its own format. Therefore, it is important for the EM information platform to be able to support storage and management of each of these different types of data.

In this section, we identify and describe the main requirements for an epidemiological information platform.

1. Supporting Heterogeneous Data
2. Enabling Searches through Metadata
3. Fostering Collaboration and Participation
4. Sharing and Protecting
5. Interfaces

2.1.1 Enabling Searches through Metadata

Since one of the objectives of an epidemiology information platform is to make it possible for users to find the data they are looking for, it is necessary it provides search services over its stored resources. The most common way to enable searches over complex heterogeneous data is to described it by metadata. This metadata must accurately describe the contents in the resource for users to be able to find it while searching. Because resources are heterogeneous, the task of characterizing and describing a resource in

metadata is complex for the standard user and, furthermore, using free text in metadata can be ambiguous.

Managing and handling metadata characterizing heterogeneous data is difficult to do in a systematic manner. On the one hand, categorization should be made on several domains. Given the heterogeneity of epidemiological resources, there must be a way to categorize them in respect to their demographic data, environmental data, biological data (e.g. symptoms and transmission modes that categorize a possible disease mentioned in the resource), etc. On the other hand, there is a plurality of synonyms, abbreviations and possible ambiguity that is not easy to manage. This is particularly true in biomedical fields, such as epidemiology, where scientific nomenclature is abundant. For example, searching for resources that refer to a class of disease, such as "viral infectious disease", can be a hard task, given that some resources may contain the phrase "viral disease" rather than the more standard name, and that other resources may not even contain the term itself and use a more specific term such as "influenza". All these resources belong to a result for a search for "viral infectious disease".

To resolve this issue, it is necessary to define a metadata model that specifies a set of elements which, when filled with the right information, describes the resource and its contents accurately. The details of the implementation of the semantification of metadata and how it is used is further discussed in Section5.

2.1.2 Fostering Collaboration and Participation

An information platform for data sharing in a scientific community thrives on user participation and collaboration. It is therefore necessary to have mechanisms that facilitate these behaviours. A typical mechanism seen in

other platforms is a way to follow the status of a resource, enabling users to keep track of the latest version of a set of data which can impact their work.

Additionally, in science it is advantageous to maintain a healthy discussion on data and existing research. Users should therefore be able to provide their insight and comment on existing resources, in a way which can lead to further collaboration or improvement of said resource.

Finally, the data a user is searching for may not always be available, therefore the information platform should give the user the possibility to create resource requests to let other users, and potential uploaders, know what he is looking for. This makes it possible for a second user to fulfil the request by creating an adequate resource, facilitating collaborative behaviour.

2.1.3 Sharing and Protecting

Resources in the information platform are uploaded by users, the owners of the resources, and therefore they must be able to share their resources with other users in order to foster collaboration. However, because resources in epidemiology may be sensitive in nature, not all resources can be shared with the same audience. It is therefore necessary to protect the privacy of sensitive resources from unauthorized access. Because data can be sensitive but not metadata it should be possible to publicly share just metadata while enabling users to privately set permissions for the data for individual users.

2.1.4 Interfaces

An information platform for storage and management of scientific data requires adequate inputs and outputs from the users and applications that access it. It must therefore provide adequate interfaces to interact with these entities. As a web based information platform, it should provide

both a browser based user interface, where the regular user can interact with the functions provided by the platform, and a web service application programming interface (API) for connectivity with external application access.

2.2 Architecture

The EM has a three tier architecture consisting of a back end, a web service and a user interface layer (see Figure 2.1) . The back-end includes repository software which stores and manages the resources, a user directory which stores all user data for authentication and a search engine to provide fast access to (sets of) resources. The end-user interface consists of web pages which provide a graphical interface, providing access to all the repository functionality by connecting to the repository via the web-services layer. The web-services expose the repository functionality to both the user interface and external applications via a web-based API.

2.3 Data and Repository

The Epidemic Marketplace deals with heterogeneous resources, such as datasets, web sites, events, software and others. The object structure of the Epidemic Marketplace should enable the search of resources without compromising the security of the data in the resources. Since EM resources are very heterogeneous, object owners should be able to organize their resources in container entities.

We propose a *folder-like* object structure as described in figure 2.2, which is similar to that seen in operative systems and therefore familiar

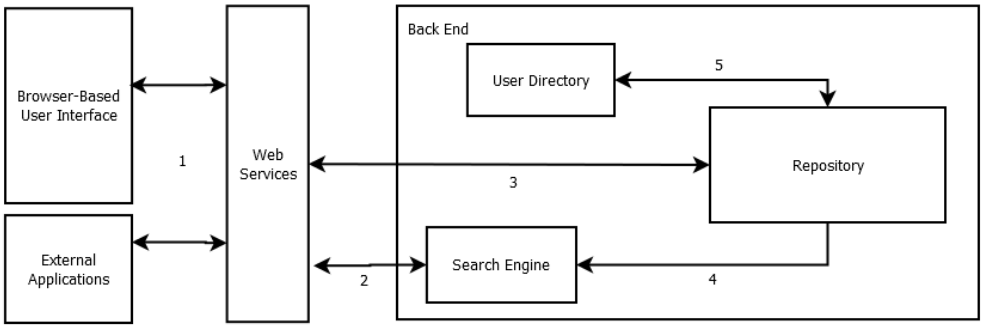


Figure 2.1: Architecture of the Epidemic Marketplace. web-service requests (1) generate search (2) and repository (3) requests. A Search Engine indexes (4) the repository. Authentication/authorization requests (5) have to be granted by the User Directory for each access.

to end-users. Resources are data content objects and also contain metadata. Collections are containers which may be composed of resources, metadata and other collections. Furthermore, this folder-like structure enables owners to share multiple resources with their collaborators without having to assign individual permissions to those resources.

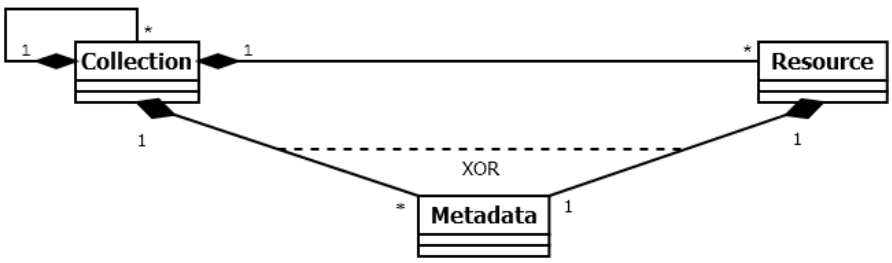


Figure 2.2: UML Class diagram of the Repository.

To implement our class model in the information platform as a software module with data storage and management facilities we explored different

repository solutions. Uploaded resources are mapped to a digital object model containing data and metadata in the repository. The functionalities implemented over this digital object model translate into the actions that can be performed over resources and collections in the EM, e.g.: create, delete, update, view a resource, etc.

Popular open source software packages for implementing digital repositories include *EPrints* [15], *DSpace* [16] and *Fedora Commons* [7]. All three options are open source software, compatible with the Opan Archives Initiative (OAI) standards (see `http://www.openarchives.org/`. These are solutions for digital repositories which include the use of metadata for the description of resources, search functionality based on these metadata, authentication and authorization functionality, as well as facilities for the storage and management of multiple types of resources.

Additionally, *EPrints* and *DSpace* are out-of-the-box solutions designed for easy installation and configuration for specific repository use cases, which do not require a technical staff to customize it, and also include the front-end user interface ready for use.

Fedora Commons, on the other end, is highly customizable in all of its functionalities, including metadata, access control mechanisms and even repository structure. However this customizability comes at the cost of requiring a technical staff to manage it. Furthermore instead of providing a repository front-end, Fedora Commons comes with an API which enables full use of all the repository's functions. The repository front end for a Fedora Commons based system must be deployed separately making use of this API.

Beacuse the EM scope is not limited to literature and textual datasets, but also software, simulations, mathematical models, web site links, etc., out-of-the-box solutions that are designed mainly for literature, such as

Eprints and DSpace, come up short in fulfilling all repository requirements. Particularly because their access control mechanisms lack customization and, therefore, does not enable the application of the required discretionary decentralized access control mechanisms. Furthermore, web services are missing from Eprints and in DSpace these services lack important functions required by the EM, such as a search web service.

Therefore, we chose Fedora Commons for digital object storage and management because it is highly customizable and can be tailored to fit the requirements for the Epidemic Marketplace. Fedora Commons provides the most complete repository software package, including Access Control services, a digital object model which can be adapted to suit the needs of a specific application and enables the creation of relationships between digital objects to implement object structure.

Fedora uses a digital object model that aggregates one or more content items into a single digital object. Content items can be of any format and can be stored locally, externally or simply referenced. This digital object model is flexible so that different kind of objects can be created although the nature of the digital object remains the same throughout Fedora Commons so that they are managed consistently.

The basic components of a Fedora Commons digital object are:

PID: A persistent, unique identifier for the object.

Object Properties: A set of system-defined descriptive properties that are necessary to manage and track the object in the repository.

Datastream(s): The element in a Fedora digital object that represents a content item.

Fedora reserves five Datastream Identifiers for its functions, *DC*, *AUDIT*, *FESLPOLICY*, *RELS-EXT* and *RELS-INT*. Of particular interest is the *RELS-EXT* Datastream, which is primarily used to provide a consistent place to describe relationships with other digital objects. The *FESLPOLICY* datastream is used to Access Control which will be detailed later in this report.

To implement our object model, we map our concepts to the Fedora Commons digital object while borrowing insight from Muradora [17], which was one of the first implementations of a repository based on Fedora Commons.

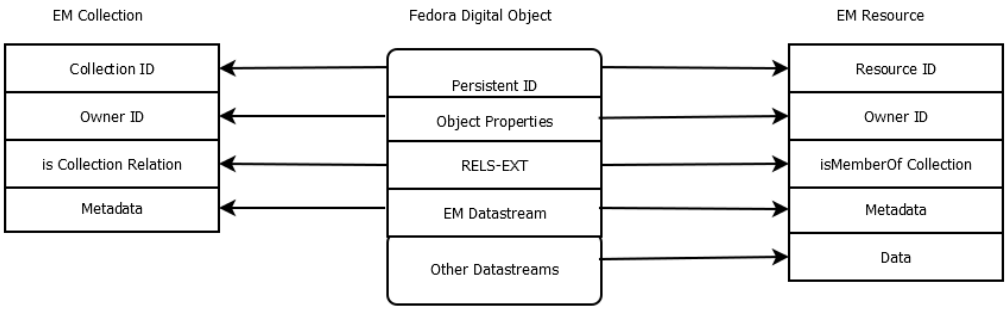


Figure 2.3: Object mapping between fedora digital objects and EM resources and collections.

A resource is mapped to a Digital Object, giving each resource a unique PID. The EM metadata extends the scope of Dublic Core elements, which are typically stored in the *DC* datastream. In our design, the metadata is stored in a reserved *EM* datastream following the EM metadata model. Resource content is stored in separate datastreams. Finally, for collections we take insight from the Muradora's object implementation and use the *RELS-EXT* datastream to map a resource to their parent collection through

an “isMemberOf” relation. Collections are composed only of the reserved datastreams, an *EM* datastream and no additional datastreams.

2.4 Access Control

ßß The EM repository stores epidemiological data, which may be sensitive. Users will only share sensitive resources in a repository if they trust it and if it provides intuitive mechanisms to control who can access them. Therefore, the EM must provide access control mechanisms to protect these sensitive data. Access control deals with authenticating the identity of users in the system and granting users the actions they are authorized to perform over resources.

The EM must provide access control mechanisms which enable resource owners to define who has access to their resources. One approach to this problem is the use of discretionary access control (DAC) [18] which enables users to define access control for their resources individually to other users. Through user groups a resource owner could assign permissions to a number of users with a single policy [19, 20, 21].

Therefore, in the EM, permissions are assigned by object owners to groups of users, granting the users access to perform an action on an object. These groups of users are also managed in a decentralized manner where the creator of the groups manages the list of users that compose it without the intervention of system administration. Additionally, permission assignment to individual users in the traditional DAC sense is also maintained in our system, and are also assigned by resource owners.

The flexibility of this type of decentralized DAC approach is that even if administrative roles are necessary for some tasks in the repository, adequate

permissions can be assigned to groups specifically created for that purpose. One such example is a group of curators, which has the task of improving the metadata annotations of public resources. In the EM, there is a user group named "Curators" which has permissions to edit the metadata of public resources, and therefore adding a user to this group gives him the necessary permissions to perform a curator role.

In our implementation, user data is stored in OpenLdap, a Lightweight Directory Access Protocol server. Users are stored as iNetOrgPerson and eduPerson class instances. This enables LDAP to store usernames, passwords, user contact information and affiliation. Authentication is performed through the Fedora Security Layer (FeSL) by querying LDAP to validate a username and password. Additionally, user groups consist of groupOfNames entities which are stored in LDAP and contain relations to one or more users registered in the system.

Authorization deals with granting users access to the actions they are authorized to perform on objects. In the EM these actions are *create*, *read*, *update* and *delete*.

Objects may be a collection, resource data or resource metadata. Permissions given at a collection level propagate to the resources contained in that collection. Because metadata is separated from data in our object structure it is possible to divulge the existence of a resource and make it searchable by making non-sensitive metadata available to the general public while protecting sensitive data with separate permissions.

Authorization is performed in FeSL using an architecture similar to the non-normative OASIS XACML architecture with a Policy Enforcement Point (PEP) and a Policy Decision Point(PDP) (see Figure 2.4). When a user requests a resource in Fedora Commons, the PEP intercepts the request and

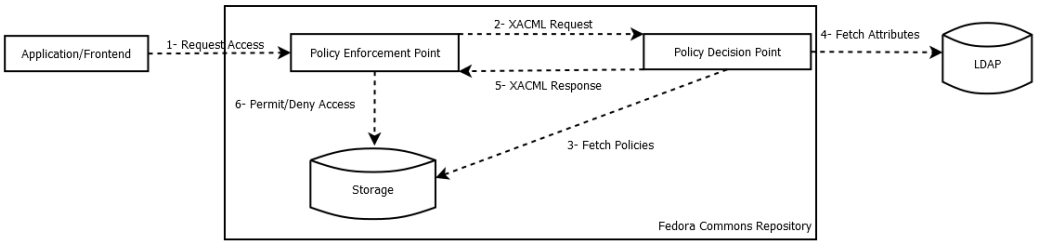


Figure 2.4: FeSL XACML authorization architecture.

forwards it to the PDP. PDP gathers the required policies and user attributes and informs the PEP on whether the user should or not be granted access. If the user has access, the PEP provides the resource.

Policies are stored in the *FESLPOLICY* datastream inside each resource. When an owner grants access to a group of users he is composing an XACML rule in the FESLPOLICY datastream. Rules are aggregated by the action and object component they give access to. According to our tests, this performs better than setting individual rules or creating individual policies for each action or each object component.

To fulfill our user group requirement, access control policies must be evaluated based on group membership. To accomplish this, we extended FeSL's PDP to retrieve group membership information from the LDAP server. Additionally, assigning policies to individual users as in standard DAC is also possible by enabling the owner to create XACML rules that target a unique username which was already supported by Fedora Commons. Therefore policies target groups or individual users, assigning them a set of actions over an object.

Because we propose decentralized management of user groups, it is possible for users to generate and be members of a large number of groups. However, LDAP is optimized to retrieve attributes of a known entity and not

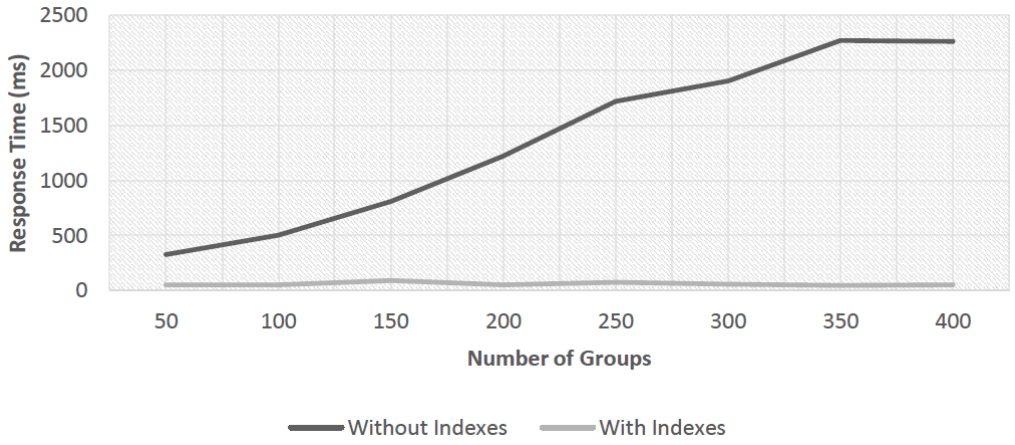


Figure 2.5: Average response time of an LDAP server with no indexes versus an LDAP server with indexes for group attributes.

all entities containing an attribute. Retrieving all groups a user is member of would mean retrieving all groups in LDAP and comparing the id of the members of the group to the user id. This does not scale as the number of groups increases. The solution involves creating LDAP indexes over group members, enabling the server to efficiently retrieve the groups that a user is member of, resulting in close to constant response times as the number of groups increases.

Figure 2.5 shows a plot comparing LDAP performance with and without indexes. The test was performed on an openLDAP 2.4 server with 51 users, 4 threads with 10 requests per thread for the varying number of groups a user is member of in a simulation where this is always a quarter of the total existing groups.

Additionally, because sometimes it is easier to create user groups based on user attributes (e.g. all the users from an institution), we enable the creation of dynamic group which are based on user LDAP attributes. To

create a dynamic group, the user specifies an attribute rule for membership (e.g. institution equals FFCUL). All users fulfilling that rule automatically become members. Unlike static groups, if a user joins the EM and he already fulfills the attribute rule, it is not necessary to add him separately to the group. Instead, he is automatically added to the group.

This grants a degree of flexibility, enabling users to assign permissions tailored to users which have specific attributes without requiring additional group maintenance over time, as users join the platform.

In our implementation, XACML policies are themselves stored in the repository. Because permission management is discretionary users can grant a large number of permissions to their groups and therefore an adequate policy storage is necessary. We tested different strategies to creating and storing policies within the repository:

1. Storing each permission assigned to a group on a separate digital object.

2. Storing all permissions for an object in a datastream inside the object with separate rules for each permission.

3. Storing all permissions for an object in a datastream and aggregate permissions into rules. One rule granting access to one action for multiple groups.

We found that the first solution, while compartmentalizing permissions, creates unnecessary workload on the repository, delaying permission management as well as the authorization process, and therefore access to resources. The first solution is therefore inadequate. When comparing the remaining two options, they performed similarly in terms of repository workload, and both had better performance than first solution. However,

the third solution improved the handling of XML, minimizing the changes that have to be made to an XACML document to grant or revoke access to a group.

Since one Epidemic Marketplace requirement is that resource owners maintain control over their resources, a generic *owner* policy grants all possible actions to resource owners. The owner is identified in a digital object property. Then group policies are created and edited by resource owners or by users with admin rights, granted by the owner, over the resource. All permissions for a resource are visible and manageable at all times by the resource owner.

Keeping authorization on the Fedora Commons repository has the limitation that SolR searches do not go through this XACML authorization engine. Because searches are simply composed of resource metadata we can assume that a user should only have access to search for resources on which he has access to view the metadata. Therefore a solution to this issue was to index the groups and usernames which have access to resource metadata together with the resource in a non readable field and filter the results of each search query to match only resources which include the groups the user is member of within our web services, presenting unauthorized accesses to resource metadata.

2.5 Resource Indexing

Because of the amount of metadata in a large resource repository, searching a repository by querying for resources that have specific metadata can lead to performance issues. A way to deal with this issue is by creating an index of metadata terms used in resources. To achieve this we used a solution

based on the Apache SolR indexing engine and FedoraGSearch, a plug-in for Fedora Commons which provides support for SolR indexing of Fedora Commons digital objects. The search web service queries SolR directly.

SolR works by using XML representations of fedora object from FedoraGSearch and indexing information from multiple object datastreams. These XML representations are sent to SolR on each resource update, guaranteeing that uploaded and updated resources are indexed according to the current content of the resource. SolR is configured to index all the metadata, which can then be queried via open text searches or over a specific metadata tag.

2.6 Web Services

One of the main requirements for Epidemic Marketplace is that it provides facilities for storage and management of data provided by other platforms, namely those in EPIWORK. In order to give access to external applications it is necessary to provide an adequate interface for that interaction. To achieve this a Web Service API was developed in Python and served in an Apache server with mod_wsgi. Connections to Fedora Common's APIs are handled by Python's fcrepo module [22], an open source community developed module that handles fedora commons API calls.

Instead of directly exposing the totality of Fedora Commons API, web services were tailored to perform the actions users can perform on the EM. Following are the web services which can be used to interact with resources and collections on the epidemic marketplace.

Create: Creates a collection or resource by uploading its respective metadata.

Upload Datastream: Creates content by adding a datastream to a resource.

List Datastreams: Lists all the datastreams in a resource.

Fetch: Retrieves a resource, with all its datastreams, or a singular datastream.

Delete: Deletes a collection, resource or datastream.

Search: Searches collections or resources indexed in SolR.

Share: Change policies for a colection or resource. Enables sharing of these entities with groups.

Add Resources to Collection: Adds resources to the collection.

List Resources: Lists the resources in a collection

Finally a set of web service methods enable the creation and management of user groups:

Create Group: Creates a group of users.

Delete Group: Deletes a group of users, but not the contained users.

List Groups: Lists all groups visible to the requesting user.

Get Group Information: Retrieves group information, including member users and visibility.

Add Member: Adds a user to a group.

Remove Member: Removes a user from a group.

Change Visibility: Changes the visibility of a group.

Web services were inspired in the RESTful principle, though some services are more complex than a fully REST approach would suggest to cope with digital library industry standards such as OAI-ORE and OAI-PMH. Further details on web service usage may be found at the EM's website documentation[1].

The search web service differ from a RESTful approach because it is implemented based on the most common practices for SolR search services using Lucene syntax.

2.7 Structures for Community Participation

In order to foster community involvement with the repository, digital objects were modified to accommodate structures for the community to manifest interest in, discuss or request resources.

Users may want to follow the evolution of a resource as it changes, as more data is added to it, or simply to keep it on a list of interesting resources to be used later. In order to enable users to do this, an additional datastream named *WATCHES* is added to digital objects. This datastream is an XML document with a list of the users which have marked the resource to be on their watch list.

```
<Watches pid="empid:111">
  <Watch username="User1" date="2013-03-20T16:16:40Z">
    User1
  </Watch>
  <Watch username="User2" date="2013-04-11T16:06:35Z">
```

[1] http://www.epimarketplace.net/developers_corner/web_services

```
    User2
  </Watch>
</Watches>
```

In the above example, User1 and User2 have set to watch the resource empid:111. Additionally, this datastream is also indexed for searches so that users can list only their watched resources.

We consider that scientific resources, such as those in the Epidemic Marketplace, benefit from active discussion by the community as it may lead to collaborative behaviour between intervening users. Keeping the discussion over the resource on the structure of the resource itself helps in keeping it in context and enables people to find the discussion while browsing the resource itself, instead of having to search for it separately. To do this we create a *COMMENTS* datastream in the resource which consists of an XML document containing a list of comments, the user who created each comment and the data and time it was created.

```
<Comments pid="empid:111">
  <Comment id="0" date="2013-04-08T13:42:25Z" username="User1">
    This resource is exactly what I was looking for.
  </Comment>
</Comments>
```

In the example above, User1 left a comment to let the owner know he appreciated the resource.

The resources being sought by a user may not be available at the repository, because they have not yet be created or because they may have been uploaded, but are private. To support the needs of users who search data that they can characterize but do not possess, we create the concept of

a *Request*. A *Request* is a digital object, similar to a resource, which does not contain any data. We created *Requests* so that they can be fulfilled by other users. To do this, we enable users to link resources to requests by adding a *isDerivationOf* relationship to the *RELS-EXT* datastream of the resource. As a result, a user can create a new resource or give the requesting user permissions to access a private resource and link it to the requests using this relationship to notify the user that he now has access to the requested resource.

A single collection contains all requests to avoid confusion between requests and resources. Additionally, objects in this collection are not listed when searching for resources. Requests consist of the metadata describing the resource the user is looking for.

Chapter 3

Availability and Distribution

The EM has been conceived as a distributed structure, with multiple nodes, each with its own repository and user directory, to be hosted in different states. In this chapter we present the hardware infrastructure on which we deployed the Lisbon node, followed by the the challenges and solutions devised for maximizing the availability and distributing the platform.

3.1 Infrastructure

The Epidemic Marketplace node in Lisbon runs on two Dell PowerEdge SC1435 servers, each with two Quad Core 1GHz CPUs, two 1TB hard drives and eight 2GB 667MHz RAM chips.

Additionally, the node uses two Iomega StorCenter 200rl network storage devices with four 1TB hard drives on a RAID 5 parity for Backups (henceforth Iomega1 and Iomega2). Backup procedures are described in section 3.2.1.

We run each componnent of the EM in a separate virtual server on top of the physical infrastructure, including the Drupal front-end, the LDAP server,

Fedora Commons, the indexing engine, web services and a sendmail server (see Figure 3.1). This enables us to manage resources more efficiently, reduces possible instability issues related to concurrently running all software on the same machine, and limits failures to a specific virtual machine, enabling recovery by the replacement of that virtual machine by a previous working version. Additionally, we can easily clone virtual machines for use in the development of the software. We use the XEN Hypervisor software for virtualization (see `http://www.xenproject.org/`).

We recently relocated all the physical infrasrtucture a newly built datacenter at the Data Center of the Faculdade de Ciências da Universidade de Lisboa (FCUL). All the EM servers are located at this infrastructure for server hosting at the institution, providing adequate housing for servers, a firewall, cooling systems which include free cooling when external temperatures are favourable, a modular UPS and power generator to maintain system availability during power failures and also a fire detection and extinction system. The data center also provides a high-speed internet connection based on optical fiber.

The FCUL Data Center, provides monitoring services through OPManager software. This software is configured to provide usage statistics of the machines it monitors. It monitors disk, RAM and CPU usage as a well as who is logged onto the machine. It is set up to issue warning on high usage of system resources and also notifies administrators via e-mail when the machine suffers from unexpected downtime.

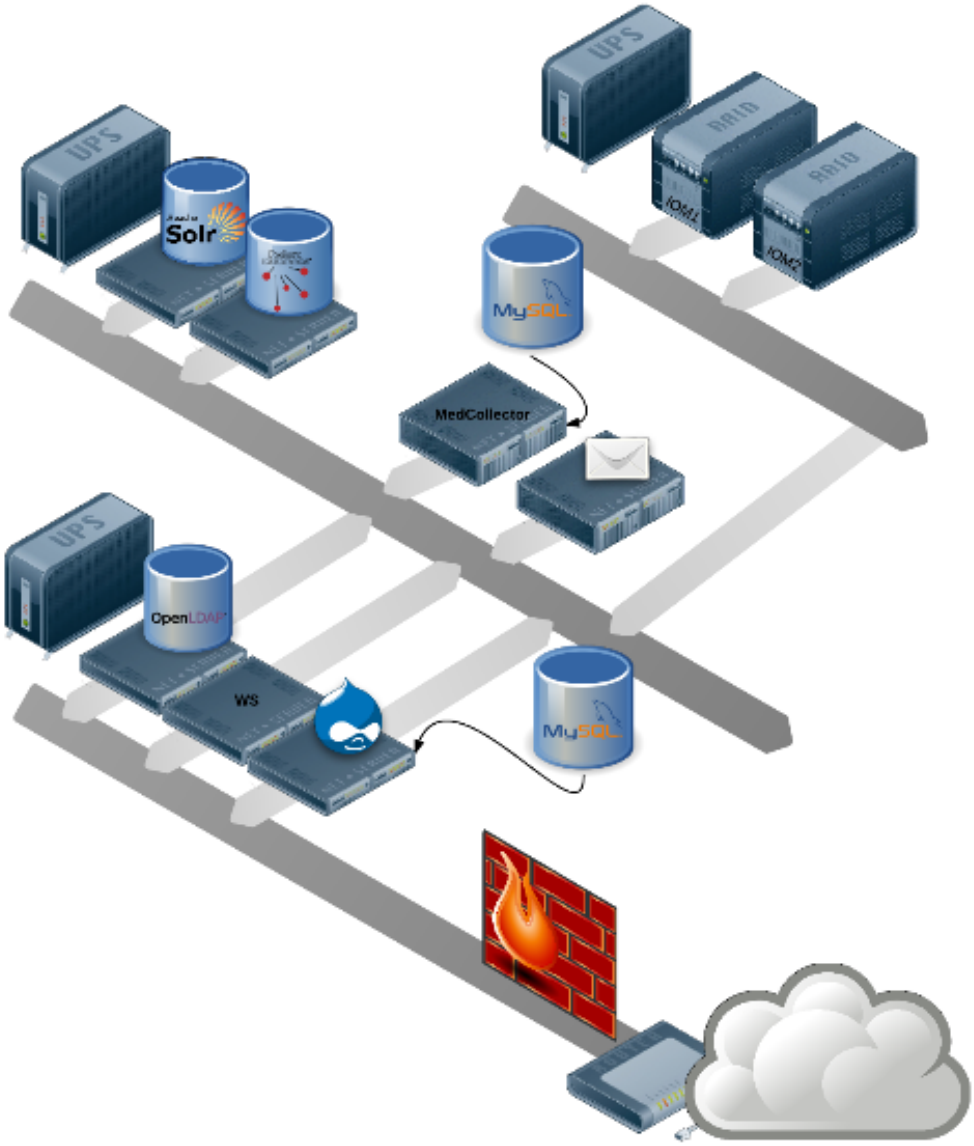


Figure 3.1: Epidemic Marketplace Node Architecture

3.2 Availability

During the development of the information platform of EPIWORK, the support infrastructure of the Epidemic Marketplace suffered several incidents, some of which severely affected its availability. Given that we were supporting our partners working remotely and the goal was to provide a service to the epidemic modellers community, we understood that we had not only to provide a modular design and stable software, but also incorporate mechnisms to maximize the platform's availability by incorporating fault tolerance. We describe such aspects of the design and implementation of the EM in this section.

3.2.1 Backup and Recovery Procedures

Backups are crucial to the recovery of failures which result in loss of data. Because of this, a series of backups are performed each day at 08:00am UMT. All backups are performed to Iomega1 and after backup completion they are replicated to Iomega2.

An image is created for each active virtual machine, both in production and development and this image is stored on Iomega1. Additionally, for each of the services access logs, configurations and production data are copied to a Iomega1 backup. Databases that support Fedora Commons, Drupal, LDAP and MEDCollector are fully backed up each saturday at 5:00am UMT, when activity at the EM platform is at its lowest, and incrementally on each of the other six days. This ensures full recovery for lost data updated less than 24 hours before a failure by the backup service.

Software packages are additionally under versioning on a subversion server at `https://epidemic-marketplace.googlecode.com/svn/trunk/`.

Having backups of every virtualized service and every virtual machine image saved as a backup and each configuration file and data also copied as a backup, failure recovery consists of getting those backups to replace the failed system.

In case of an unrecoverable failure of a virtual machine running a service, the backup image is used to re-create the virtual machine to replace the existing failed machine. In the case of failure of Iomega1, the backups are replicated on Iomega2 to avoid backup loss. If a virtual machine host server fails, virtual machine images can be used to deploy all the services on a different virtual machine host server. These failures can be recovered from in a short period of time.

In case of virtual machine image corruption causing the inability to recreate the virtual machine as it was prior to failure, because all configuration files and data are also under backup procedures, it is possible to create a new virtual machine, deploy the versioned software on the machine and deploy the backups configurations and data. This requires additional downtime as it would require more administrative intervention than the previous cases but it still guarantees full failure recovery of data and systems to the state at the time of backup.

3.3 Distributing the EM

This section covers the details on the distribution of the EM platform across several geographically dispersed, fully functional EM platform nodes, the EM Nodes.

This distribution is a result of the requirement that resources have to be stored separately due to privacy concerns. Several resources can contain

more sensitive data, and thus requires being stored in a separate repository, managed by a specific owner under the laws of a specific country.

Each node implements a set of services, including:

- The front-end interface
- A Fedora Commons repository
- An indexing Engine
- The Web Services
- A LDAP authentication mechanism
- A Mail service
- The MEDCollector

Each service can be implemented in a separate server. This separation allows for service replication and redundancy, providing fault tolerance. It is also possible to have more than one running instance of each service at the same time, provided they are configured and integrated to work as such. We have tested such environment, and concluded it helps further improving each node's availability.

Since each node provides all services, it implements a fully functional EM platform, with its own resources, and capability to respond to services, which allows for requests to be dispersed throughout all nodes, with each client accessing the functional node closest to it. This distributed scheme will allow better load distribution and scalability. In addition, the level of distribution may also help in implementing distinct access policies complying with different legal requirements for protecting data in various states. For instance, there may be a legal requirement for storing a dataset in a specific

country, while the metadata of the dataset could be replicated in multiple EM nodes.

3.3.1 Connecting Multiple Nodes

In this section we cover how each node can communicate with each other to enable resource or metadata access from each single node, while dealing with the privacy requirements of sensitive data. Because data creators are responsible for the managment of the permissions over their resources. We present three approaches to solve the this problem and which we think is the more adequate solution.

The first (Figure 3.2) consists of a centralized Fedora Commons repository containing all the actual resources stored in other EM nodes. This is done by storing all the metadata in the centralized repository and having metadata that links to either internal or external datastreams containing the actual data. Instead of a decentralized solution, this consists of a centralized repository with a distributed file system.

While this solution requires less maintenance due to it only requiring a single repository, it lacks in decentralization and requires all access control to be managed on the central repository, which can be an issue when dealing with nodes that may have different data privacy requirements.

The second solution (Figure 3.3) consists of deplying multiple Fedora Commons instances, each with their own resources (which may or not be replicated), and using Apache SolR servers to index resources at each node. Each of the nodes has an Apache SolR instance that indexes all the resources in the node, meaning all servers must be queried in order to obtain the results from all the existing nodes. This approach effectively decentralizes access control, enabling each node to exercise their access control mechanisms

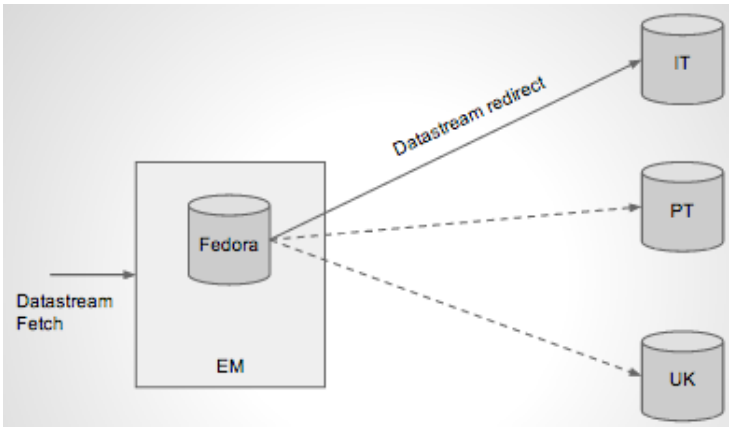


Figure 3.2: Centralized Fedora repository with distributed file system.

independently. The disadvantage of this solution is the performance overhead of communication between several external nodes to get a single resource. If a query must be made to several servers it may take a considerable amount of time to be answered.

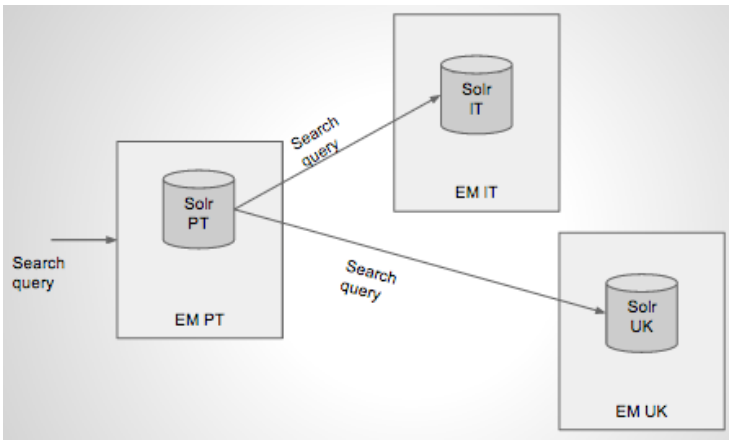


Figure 3.3: Apache SolR servers across several countries resolve a query.

A third, final approach (Figure 3.4), consists in having the Apache SolR instances synchronized, including the location of the resource. Each Apache

SolR instance will periodically send the additions to its index to the other SolR instances. A query to either one of the indexing engines wouldn't have to be propagated to respond to a query. While there is a delay in this synchronization mechanisms, we still feel that it is the best solution to connecting multiple nodes because it does not affect service performance and still maintains decentralized access control mechanisms on each node.

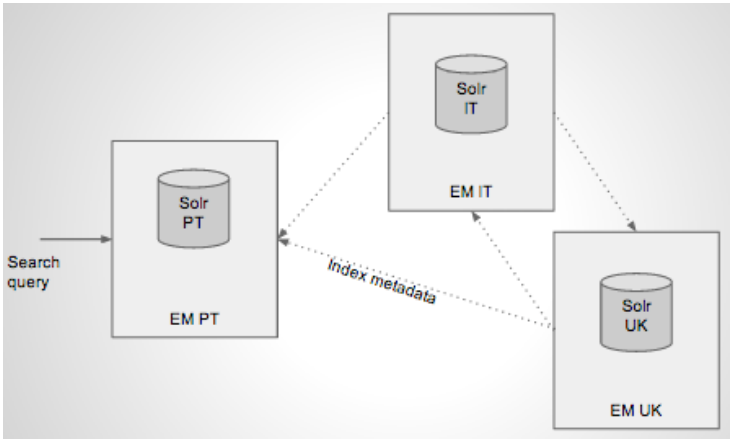


Figure 3.4: Replication of SolR indexes.

Chapter 4

User Interface

This chapter reflects on the main design choices for the user interface of the Epidemic Marketplace. Our primary concerns in the creation of the interface for the EM were:

- The Development of a Data Organisation and Categorization Model under a consistent and coherent informational architecture for the Epidemic Marketplace.
- The Creation of a faceted interactive interface for the Epidemic Marketplace.
- The embodiment of the graphic interface on the Epidemic Marketplace system.

Due to the large amount of data we face everyday, it is impossible to understand or visualize all the represented objects, their categories, even relationships in it. Currently, the computing power allows to get data as never before. It actually allows to get data on data - metadata - which in its turn, is disseminated through web protocols, such as RSS feesd. These protocols turn data easily accessible, manipulable and redundant.

"Information organization and visualization can be, therefore, defined as the use of computer supported, interactive, visual representations of abstract data to amplify cognition, discovering and highlighting relationships in data elements. A proper visualization, after a proper selection and filtering is actually a kind of narrative, providing a clear answer to a question without extra details. It simplifies human comprehension." (Nathan Shedroff) [23]

4.1 Design Process

"To lay out paths, first place goals at natural points of interest. Then connect the goals to one another to form the paths." – Christopher Alexander

We centred or design process on this approach. First and based on our informational structure we identified the content goals, we draw previsible inward and outward paths between those content goals. After that we went to a process of paper prototyping, and built sketches of the main pages including the homepage. We draw vectorized mockups based on the paper sketches and finally completed the web implementation.

Based on Christopher Alexander's approach, we identified five content goals which are the primary points of interest for the user. They are:

Resource: our most important asset. The aggregated data and metadata.

Request: a data requirement posted by an user.

Announcement: a type of message targeted for a specific user or the general public.

Content page: a type of informational content that we couldn't put in a specific category.

Figure 4.1: Homepage mockup final proposal

Web service: this type of content is targeted for developers and it's main purpose is to help them using the Epidemic Marketplace Application Programming Interface (API).

The paths between content goals

Based on Christopher Alexander's approach we defined content goals and joined them between paths. These paths are the way the information flow from one page to another, or the way in user drills for more information that can be related.

The diagram 4.2 represents the content map of this site, the Goals and how they relate to each other. The map was based on the previous application releases but with an user focus perspective.

Paper prototyping

Paper prototyping 4.3 was the easiest way to explore ideas on how could we draw the interface, without being concerned on details at that phase. It was revealed as a simple way to communicate and share ideas.

Vectorized prototyping

We wanted our design to be simple, clear and objective. Given the focus on the user experience and his needs, to be user-centred. We draw vector pages for every content goal and listing and pages. This vectorized pages were still a mockup of the final design, but they aided in making the vision for the final form of the design clearer without the need to yet concern for the color palette.

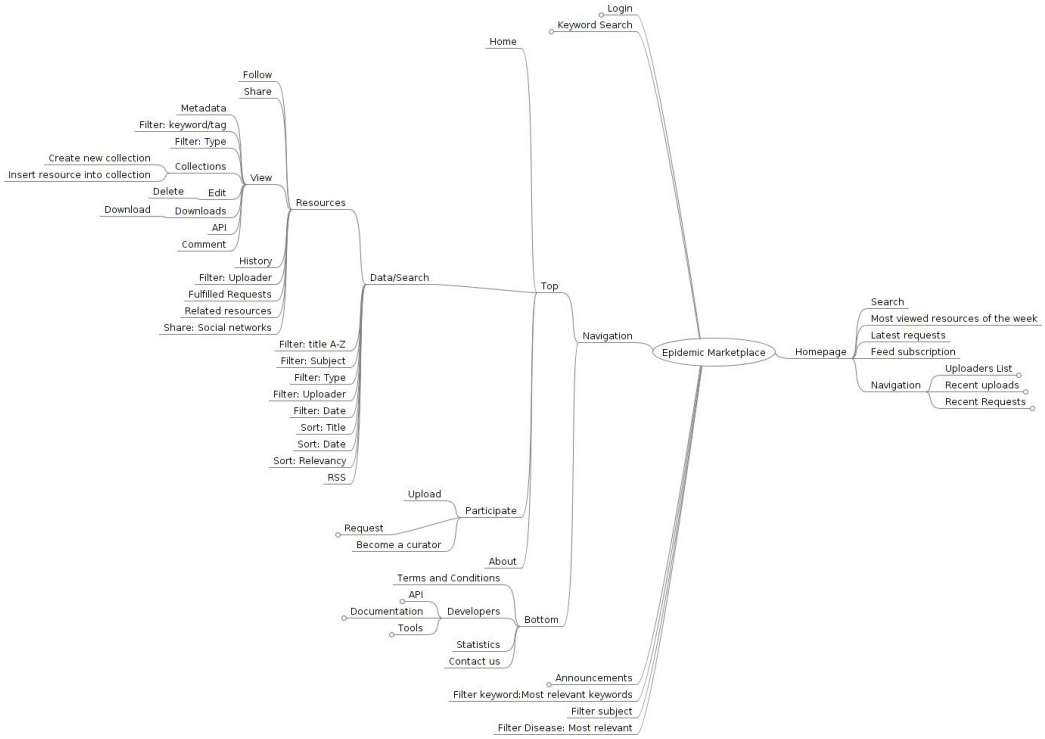


Figure 4.2: Map that represents the architectural structure of the site. Notice how most of the interaction is centered on how to access the previously defined content goals.

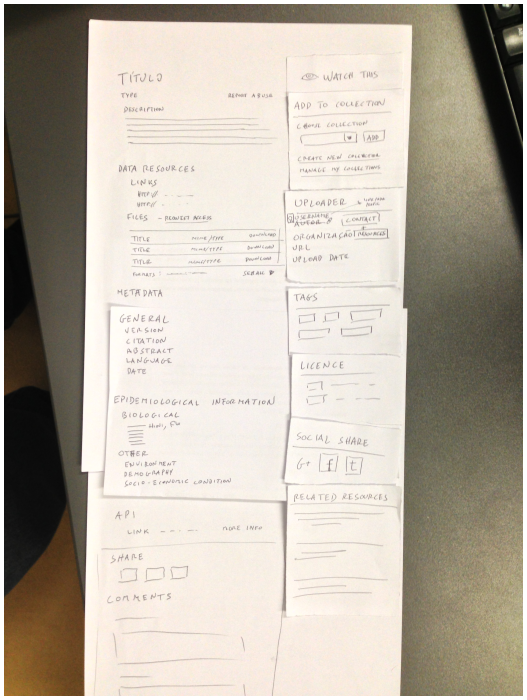


Figure 4.3: Paper prototyping stage of the Resource Page

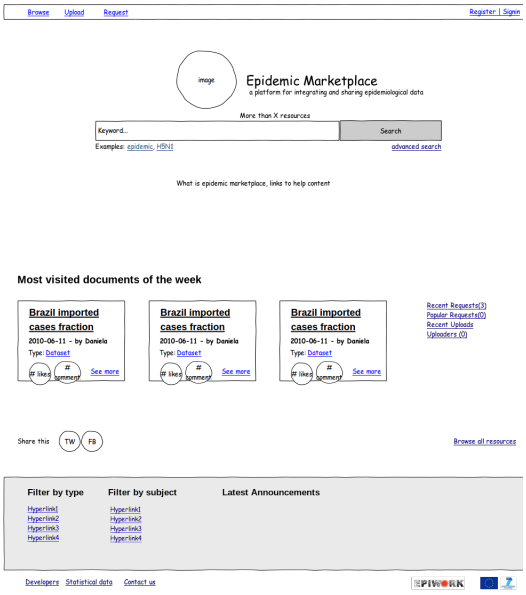


Figure 4.4: Vectorized prototype stage of the home page

4.2 User Interface Design

The final design had several influences:

- The color palette of EPIWORK project;
- The way Youtube's videos results were presented when searching;
- Other Open Data websites, such as `http://data.gov.uk`.

While creating an information or graphic architecture it is mandatory to have in consideration elements such as main colors; font types and their weight, size or color; text format and the position and sizes of the structural elements along the page. On the Epidemic Marketplace, a graphic identity was the key requirement and had to be created so as to give some more visual consistency. That consistency is crucial for the user to relate all the platform's different modules and do its main operations. So the first step in

the graphic process was to create a logo to identify the project and a general Style Sheet 4.6. This Style Sheet presents the style for major aspect in the website. Things like Headings, text, boxes, form fields, colour's palette, iconography.

Figure 4.5: Epidemic Marketplace logotype

The Epidemic Marketplace design has been an evolving task. To accomplish the current layout we have tested previous versions of the website through online graphical user interface evaluation services to spot potencial problems. Figure 4.8 shows how the previous layout drew attention away from the main functions of the site (browse, search, request and upload).

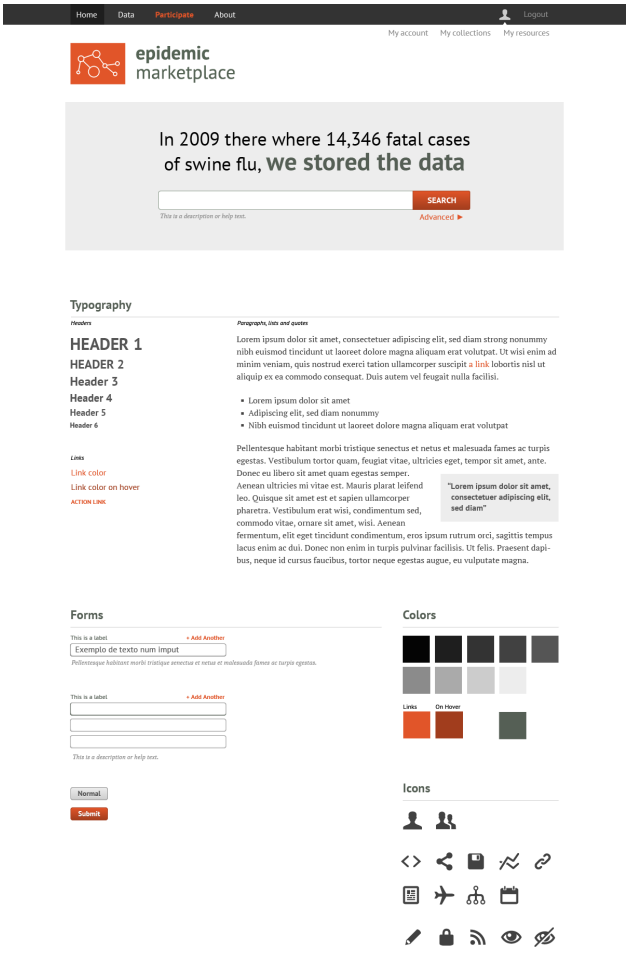


Figure 4.6: Epidemic Marketplace stylesheet

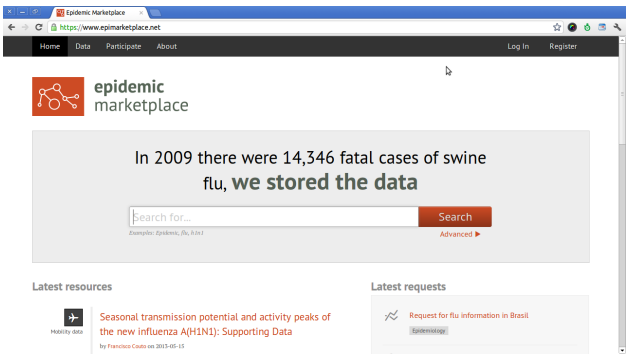


Figure 4.7: Homepage screenshot

Figure 4.8: Older homepage tested against Feng-GUI.com service

Resources are our most important asset. It's an abstract entity that represents one epidemiological resource. It's also composed by several datastreams, one of them is his metadata.

Requests , like resources, are abstract entities. They represent a requirement of epidemiological data posted by an user. It's also composed by several datastreams, one of them is his metadata.

Announcements are clear and small messages for users. They are composed by a simple message, author, date and time and also a recipient. This recipient could be any registered user or the general public. In this case the message is public for anyone on the Internet.

Web services interface assets This content type is targeted for developers. It's main purpose it's to help them as much as possible using the EM API. We tried to describe as much as possible each service, given usage examples to led developers participate in the process of exploring the Data through the EM API.

Chapter 5

Metadata and Ontologies

Epidemiology is data-driven and as such its models of disease spread rely highly on large amounts of information. This information can either be very general and easily located (e.g. the size of a population), or specific to a certain population and disease (e.g. contact and transmission networks). Finding highly specific information in literature is difficult, when not impossible, particularly in the time frame where those models are useful. The modelling and monitorization of real time disease outbreaks requires the quick collection of necessary information. The EM aims to address this issue by supporting data sharing mechanisms. To do this it is necessary to standardize as much as possible the sharing of digital resources. The EM establishes a metadata model for annotating resources. The metadata model provides a structure for describing a resource and also specifies the valid values that can be used in the metadata characterization of an epidemiological resource.

5.1 EM Metadata Model

The EM uses a metadata model that enables a tractable management of its resources, easing the sharing between the people that benefit from the information it stores, such as public health scientists. This metadata contains: (i) *technical information* (the uploader, an internal identifier and the date of submission); (ii) *general information* related to the digital resource, (*e.g.* title, creator – which need not be the same as the uploader); and (iii) *content-specific information*, such as the subject, the sources used by the resource or even the epidemiological information that makes up the resource.

The use of unique resource identifiers for EM resources ensures that they can be uniquely and persistently identified and retrieved in the future.

The metadata of the EM serves primarily these objectives:

- managing and organizing data contained in the EM;
- exposing the metadata to other communities, enhancing data sharing and information finding;
- being an integrating and exchanging format, flexible enough to be used with different metadata standards.

The EM needs a general schema to describe the extremely variable data it harbors. Given its success in semantic web, we decided to base the EM metadata model on the fundamentals of the Dublin Core (DC) and the DC Metadata Initiative (DCMI) Metadata Terms.

The terms offered by the DC can already handle many of the requirements of the EM, especially in the *technical* and *general* information areas. However, epidemiology relies on multiple domains of knowledge. Accordingly,

the metadata model devised for this purpose must extend the core elements of DC with tags appropriate for these domains of epidemiology. For example, many epidemiological resources deal with one or more diseases, a concept absent from DC; as such, the EM metadata model contains a specific element, `<em:disease>`, suitable for annotating a resource with the diseases it refers to. This ensures that the resource is searchable not only based on the general information provided by the DC but also based on its epidemiology-specific contents.

Furthermore, we decided to refine some of the existing DC elements. For example, the element used to annotate the creator of a resource must be filled with the name, organization and URL of the creator. As such, the creator tag is refined by three tags, one for each of those values (see Figure 5.1). The content-specific element `<em:biologicalInformation>` is also refined by a number of biologically relevant elements, such as the previously mentioned `<em:disease>`.

For consistency, it has been stipulated by the EM developing team that the elements of the metadata model should be grammatical names; for this reason, we have renamed the elements `<dc:spatial>` and `<dc:temporal>` to `<em:location>` and `<em:time>` respectively. Under the same principle, instead of using the actual DC terms, we created homonym tags in the "em:" namespace. Thus, `<dc:creator>` changes to `<em:creator>` in the EM metadata model. Additionally, the metadata model specifies the expected values that can be used to fill each element. Some expect literal values, such as `<em:title>` and `<em:date>`, which expect a string and a date, respectively. Most of the content-specific information must be included from ontologies of an appropriate domain. For example, to fill the `<em:disease>` element, instead of the literal "flu", the URI `http://purl.obolibrary.org/`

```
em:creator
    em:name         : M. Evans
    em:organisation : University of Lisbon
    em:URL          : http://mevans.com/
```

Figure 5.1: A snippet of the hypothetical annotation of a EM resource created by someone named M. Evans.

`obo/DOID_8469` should be used. This is the identifier of the concept named "Influenza" in the Human Disease Ontology. Several ontologies have been collected in a network of relevant ontologies named NERO (see section 5.2), which have been integrated in the EM so that users of the platform can search them and correctly annotate their resources. Table 5.1 contains the EM metadata model and its mapping to DC.

5.2 NERO

Most of the *specific-content* elements of the EM metadata model are filled with concepts from ontologies. However, it is unreasonable to expect all users to be familiar with the ontologies of the biomedical domain, which terms they contain and their identifier. Therefore, we integrated into the EM a number of ontologies that provide appropriate concepts, which can be searched by name and synonyms, enabling its use during the annotation process. These were collected into a Network of Epidemiology-Related Ontologies (NERO) [11].

The ontologies contained in NERO were selected based on a number of requirements [11], some of which are related to the preservation of epidemiological resources. For example, these ontologies are required to

Table 5.1: The current version of the EM metadata model specifies these elements. Many of them are mapped to the homonym DC counterpart

Technical information	**Generic-content information**
em:identifier[1]	em:title[1]
em:dateSubmitted[1]	em:generalDescription
em:uploader[2]	em:abstract[1]
em:uploaderName	em:citation
em:uploaderOrganisation	em:description[1]
em:uploaderURL	em:DOI
em:uploader	em:format[1]
	em:ISBN
Specific-content information	em:language[1]
em:biologicalInformation	em:pubmedID
em:diagnosticMethod	em:subject[1]
em:disease	em:type[1]
em:symptom	em:URL
em:drug	em:version
em:host	em:date[1]
em:pathogen	em:creator[1]
em:transmission	em:creatorName
em:vaccine	em:creatorOrganisation
em:vector	em:creatorURL
em:environment	em:organisation
em:demography	em:organisationName
em:socioEconomicConditions	em:organisationURL
em:location[3]	em:source[1]
em:time[4]	em:sourceName
em:from	em:sourceURL
em:to	em:sourceDescription
em:moment	em:bibliographicCitation[1]
	em:refCitation
	em:refDOI
Mappings:	em:refPubmedID

provide textual definitions for their concepts, to be popular among the communities that use them and to be publicly available. All these properties contribute to the preservation of metadata integrity.

In our search for the right ontologies for NERO, we found the BioCaster Global Health Monitor [24], "an early warning monitoring station for epidemic and environmental diseases", which is based on an ontology. We also found the Human Genome Epidemiology Network (HuGENet) [25] and the Dictionary of Epidemiology [26], which are not ontologies but thesauri. Overall, these have a low coverage in domains such as geography or diagnostic methods, are not as thorough as needed for the purpose of annotation, and do not seem to offer many guarantees of future support. Nevertheless, they provided a sense of the concepts that should be modeled in an epidemiological resource.

There are other general-purpose resources which contain, among others, concepts of epidemiological interest. These include the Unified Medical Language System (UMLS) [27], and the Medical Subject Headings (MeSH) [28]. From a preservation point of view, these resources are adequate for annotation. However, properly scanning through these large terminologies and determining which of their concepts are relevant would be too colossal a task for the typical epidemiologist.

In face of these issues, we turned single-domain ontologies in the biomedical field. The Open Biomedical and Biological Ontologies (OBO) is a project run by the OBO Foundry that aims to provide a suite of orthogonal, interoperable, reference ontologies in the biomedical domain [29]. It defines a set of principles that must be fulfilled by an ontology before it can be included, enforcing good quality by promoting good practices in ontology development. In particular, these ontologies are public domain

and must guarantee versioning, documentation, etc., which also contribute to manageable preservation of their contents. In addition to their high quality, OBO ontologies span over many biological and biomedical domains of knowledge, and given their association to a high profile initiative, are more likely to be kept available and up-to-date in the future. As such, we included several OBO ontologies in NERO.

Concepts from domains which are not biological must be retrieved from other resources. Yahoo! GeoPlanet^{TM}contains a representation of the world geography, and is a very good candidate for inclusion in NERO. We have been unable to find ontologies that specifically represent demography or social and economic conditions, and suspect that none exist that are publicly available. As such, NERO relies on MeSH and the Epidemiology Ontology in these cases.

Table 5.2 maps the EM metadata model elements into the NERO ontologies that provide the concepts to fill them.

5.3 NERO and the Epidemic Marketplace Metadata Model

The conjugation of the EM metadata model and NERO ontologies provides a simple but efficient way to describe epidemiological resources under a shared semantic structure. Epidemiology researchers can provide semantic annotations in several key areas modeled by the EM metadata model, and most of the specific-content elements of the EM metadata model are designed to be filled-in with concepts from ontologies integrated into NERO.

When a user uploads a resource to the EM, she is required to fill-in a few mandatory metadata fields, including `<em:title>` and `<em:description>`.

Table 5.2: Mapping of the metadata elements of the Epidemic Marketplace metadata model into ontologies that contain concepts useful to describe epidemiological resources. Each element can be mapped to more than one ontology, which is useful when neither covers 100% of the domain in question.

Metadata element	**NERO ontologies**	**Provenance**
`<em:diagnosticMethod>`	NCI Thesaurus	OBO
`<em:disease>`	DOID	OBO
	IDO	OBO
`<em:drug>`	ChEBI	OBO
`<em:symptom>`	SYMP	OBO
	HP	OBO
`<em:host>`	NCBI Taxonomy	OBO
`<em:pathogen>`	NCBI Taxonomy	OBO
`<em:vector>`	NCBI Taxonomy	OBO
`<em:transmission>`	TRANS	OBO
`<em:vaccine>`	VO	OBO
`<em:environment>`	ENVO	OBO
`<em:location>`	GeoPlanet™	Yahoo!
`<em:demography>`	Branches of MeSH	NLM/NIH
	Epidemiology Ontology	HuGENet
`<em:socioEconomicCondition>`	Branches of MeSH	NLM/NIH
	Epidemiology Ontology	HuGENet

However, the user can also provide a detailed epidemiological description of the resource by filling in the other metadata. To aid in this task, the EM provides the previously mentioned search functionality, backed up by an autocomplete function that returns concept suggestions retrieved from the NERO ontologies appropriate for the element in question. This effectively hides the technical details of the ontologies from the regular users, letting them focus on semantic annotation.

Each NERO concept is associated with its description, which the user can read to help her choose the concept that better describes the resource. Whenever a given characteristic of the resource cannot be accurately described by any of the available ontology concepts, the user can easily assign a more general concept, which is supported by the inherent hierarchical nature of ontologies. Furthermore, if desired, the user can also submit a request for the addition of the more specific concept.

Given the variable nature of epidemiological resources, not all resources will need to be annotated in all metadata fields. For instance, a resource focused on tracking the geographical spread of a disease probably won't refer to any drugs, or if it focuses on the treatment of a disease, it might not include information on diagnostic method. In a recent analysis we conducted of semantically annotating over 100 Epidemiological resources in the EM, and found that all resources mentioned at least one disease and one geographical location, about 80% included information about the diagnostic method and the pathogen involved, but only about 30% mentioned any drugs or vaccines.

A crucial feature of the EM and NERO integration is the ability to assign multiple concepts to the same metadata field, since many resources mention multiple diseases, symptoms, drugs, etc., mirroring the wide scope of epidemiology. This effectively enables crossing information from different

resources referring the same or similar entities, such as diseases or drugs, to support broader studies.

The adoption of a metadata model to support the semantic annotation of epidemiological resources, ensures a more structured annotation process, effectively guiding the annotation itself. Furthermore, by coupling the metadata model with NERO, the annotation process is further simplified, since terms to fill-in metadata fields are retrieved from a controlled vocabulary which is backed by the rich properties of ontologies such as hierarchical structure, definitions and other properties and relations.

Chapter 6

Data Sources

This chapter describes the strategies we developed for populating the data repository and promote its use. We begin by exploring the integration work done with the GleamViz simulator (the computational platform of EPIWORK) to foster the storage of simulations on the EM, and exchange of users between the two platforms. We then explore the strategis for collecting and storing anonymized Influenzanet (Epiwork WP5, `http://www.influenzanet.eu`) data on the EM, collected from the central Influenzanet database. Finally we describe MEDCollector, a system for collecting data from multiple heterogeneous online data sources and packing it into datasets of relevance to the Epidemic Marketplace.

6.1 GleamViz

This section covers the integration plan of the GLEAMviz epidemic simulation platform with the Epidemic Marketplace information platform. The integration of these two platforms provides GLEAMviz with increased simulation and visualization sharing capabilities and provides a means to

store old simulations without increasing the storage needs of the GLEAMviz Simulator. The integration with GLEAMviz also provides the EM with a new source for fresh epidemic datasets which result from simulations and also has the potential to increase its user-base with people which access the EM's repository through the GLEAMviz simulator. The EM was designed as the information platform service for use with the remaining EPIWORK platforms and by integrating GleamViz it achieves that goal with computational platform.

GLEAMviz is designed to let users define epidemic spreading simulation, execute them on a dedicated server, and retrieve and analyze the output data. Simulation definitions are described by an XML document while simulation's output data (for successfully executed simulations) are constituted by a set of custom-format data files, one for each day of the computed simulation. One of the main goals of the integration between the GLEAMviz simulator and the EM is to exploit the storage and sharing facilities offered by the EM repository to let users organize their simulations, share them with other users, and simplify team-working.

Given this scenario it is necessary to establish a proper mapping of GLEAMviz simulation's definitions and simulation's output data sets with EM objects.

A GLEAMviz simulation is associated to a resource in the EM. A simulation definition is stored as an *definition* datastream within the resource, and simulation output data is mapped to a list of datastreams each corresponding to the output of a single day. Information describing the simulation is mapped to resource metadata.

This distinction between simulation definition and simulations output data is fundamental for the standard GLEAMviz users, allowing them to

flexibly share only the first, only the second or both with other users.

The interaction between GLEaMviz and the EM is depicted in Figure 6.1

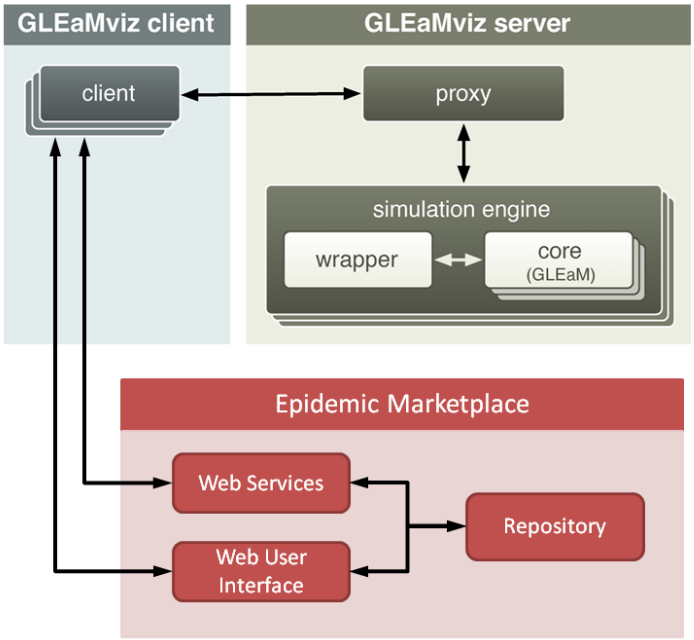


Figure 6.1: General architecture of the GLEAMviz interaction with the Epidemic Marketplace.

The GLEAMviz client (GC) software uses the EM web service API to upload, read, edit, delete, search and download simulations in the Epidemic Marketplace. Access control is managed in the EM browser interface, so that a simulation shared on the EM platform can be seen by the receiving user on his GLEAMviz client.

GLEAMviz Simulator (GS) and the EM users are authenticated across a synchronized infrastructure enabling GC and EM users to transparently access both platforms with the same credentials. This also enables the GC client to avoid additional user-credential requests upon interaction with the

EM repository, therefore allowing seamless single sign-on access to both platforms. This synchronization helps foster participation of EM users in the GLEAMviz platform and vice-versa.

GS users are be organized in two organizational units, one of them private (not shared with EM), and a set of users which are shared with EM and can access both platforms. EM users are shared with GS for the same purpose, but without the need for a private organizational unit (see Fig. 6.2). To avoid username collisions, users are identified as belonging to GLEAMviz or EM through assignment to separate Organizational Units.

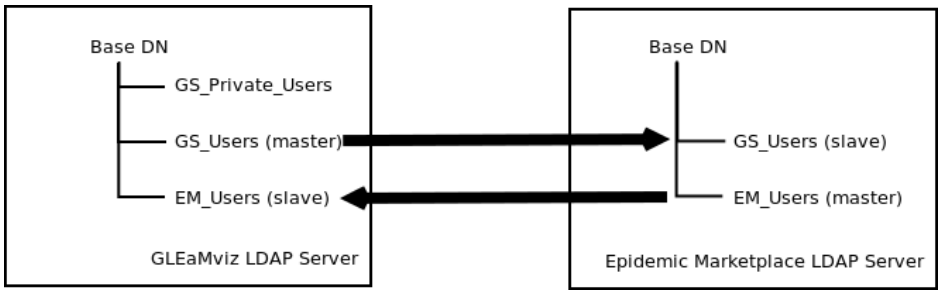


Figure 6.2: Schematic of the synchronization mechanisms between authentication servers in GLEAMviz and the EM.

Both platforms use the OpenLDAP software to manage their users and will use a multiple master configuration for synchronizing their users. Considering Figure 6.2, the GS is the master, or provider, for the Organizational Unit "GS_Users" and EM is the slave, or consumer. Conversely the EM is the master, or provider, for Organizational Unit "EM_Users" while GS is the slave or consumer. This is achieved by setting two OpenLDAP syncrepl instances. Furthermore having a replicated, synchronized, copy of user credentials, as opposed to a simple query based approach, enables EM users to use GLEAMviz even if EM is not available

during a maintenance period, or vice-versa.

6.2 Influenzanet

Work package 5 is responsible for the ICT monitoring system, Influenzanet. It has been operational on the Netherlands, Belgium, Portugal, Italy, the United Kingdom, France, Spain, and Sweden. Influenzanet generates data from weekly user surveys about the presence of flu symptoms, which contracts with the traditional system of sentinel networks of primary care physicians. As a result, data on ILI rates is generated faster which can then be compared with the real incidences in those countries.

This direct user-centered monitorization generates a good amount of data which can be of use in further epidemiological research. However, data collected from users should be adequately anonymized before it is shared on a large scale. To enable data sharing in the Epidemic Marketplace we developed a solution based on an interface with the central Influenzanet database (see Figure 6.3).

Data from each Influenzanet country is anonymized and the questionnaire answers are stored in the central database. A query is run periodically on influenzanet's central database to generatea dataset with collected data from the period. These data are moved to the EM to be sliced into a set of resources on the information platform, one for each country for that period of time.

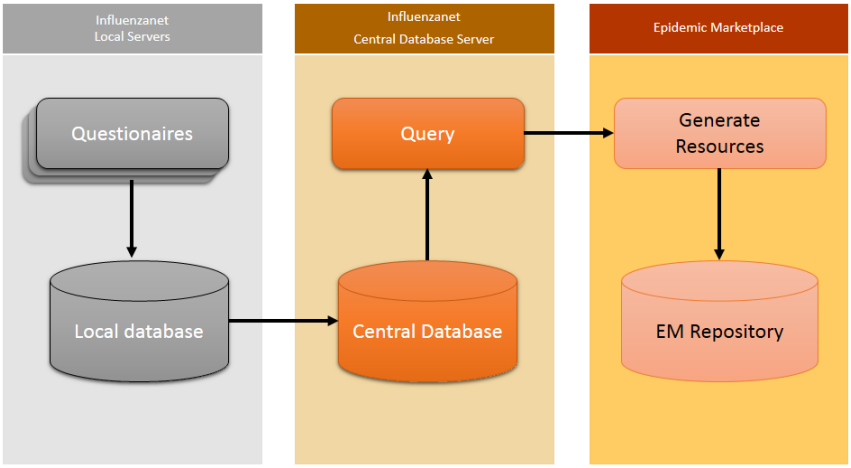


Figure 6.3: Process from Influenzanet data collection to Storage in the Epidemic Marketplace.

6.3 MEDCollector

Devising mechanisms to retrieve information from multiple web based sources can be complex as formats, types of data, retrieval methods and even periodicity change depending on the source of the data. In a typical scenario, collecting and integrating data from multiple sources in the epidemic marketplace would require that separate mechanisms would be developed for each of the different data sources.

MEDCollector is a web based software which enables users to create mechanisms for data retrieval from multiple web based sources by designing work flows based on the combination of existing web services. By enabling flexible design of workflows, MEDCollector enables users to tailor data retrieval mechanisms to each of the data sources, facilitating data integration.

Additionally, MEDCollector also enables similar workflows to be created for packaging the collected data and its upload to the EM platform.

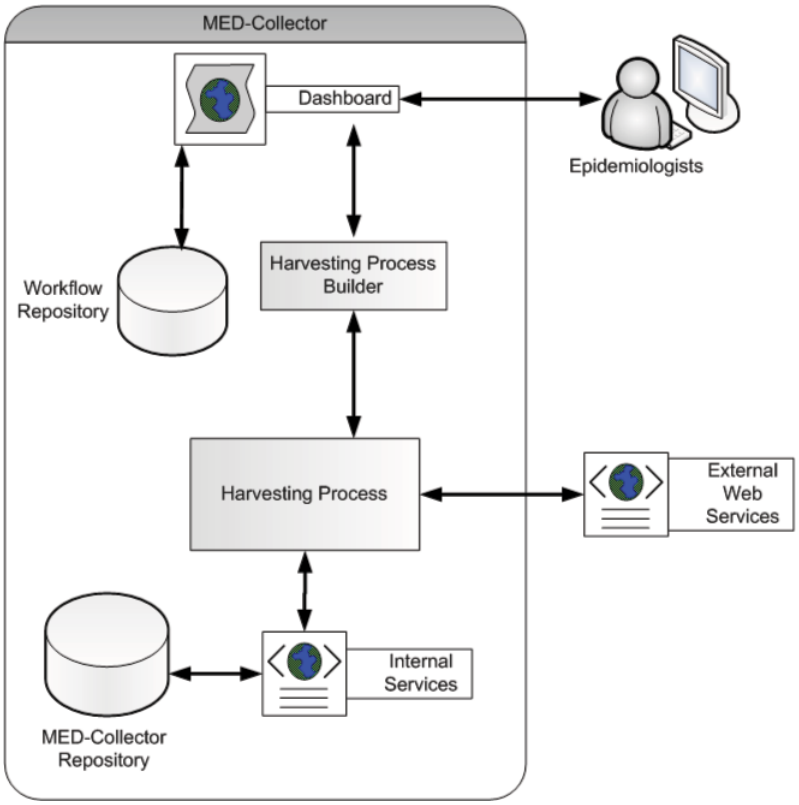


Figure 6.4: MEDCollector's software architecture.

The architecture of the MEDCollector is represented in Fig. 6.4. Its main system components are:

- *Dashboard.* Provides user-interface capabilities to the system, enabling the user to define harvesting processes and to monitor currently deployed processes.
- *Workflow Repository.* Stores workflows designed in the Dashboard.
- *Harvesting Process Builder.* Converts designed workflows to deployed Harvesting processes.
- *Harvesting Processes.* Processes that orchestrate communications

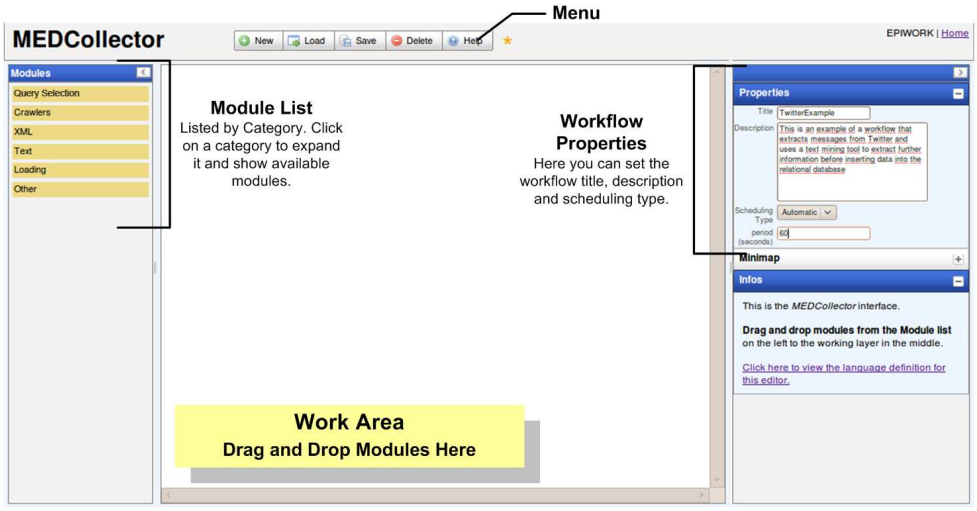


Figure 6.5: Global view of the Web Interface implemented using WiringEditor and description of its components.

between multiple services, both internal and external, to perform data collection from external sources accordingly to workflow definition.

- *Internal Services.* Provide basic system functionalities and interact with the MEDCollector Repository.

- *MEDCollector Repository.* Stores all the data collected by the system and consists of a relational database.

MEDCollector provides a graphical interface for the design of BPEL processes which orchestrate web services to collect data.

It is also bundled with a single-page editor that enables the definition of workflows through a wirable interface, see Fig. 6.5. The wirable interface consists of drag-and-drop elements which can be connected with wires between their inputs and outputs. Workflows designed in this interface are saved to the Workflow Repository.

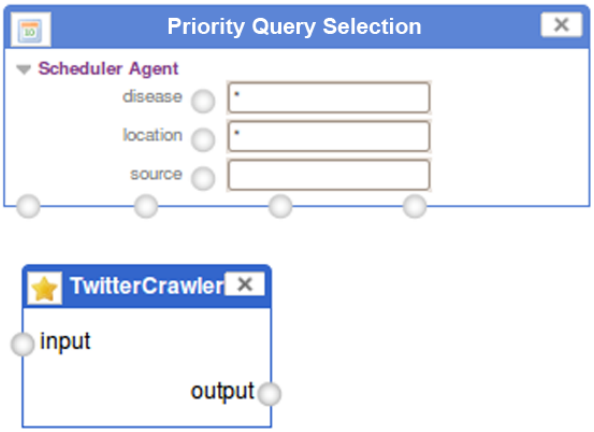


Figure 6.6: Start by adding a Scheduler and an Harvesting Service to the Work Area.

6.3.1 MEDCollector Example

This section illustrates the creation of workflows in MEDCollector, using again the extraction of messages from the Twitter Social Network as an example. Fig. 6.6 through to 6.10 presents a step-by-step description on how to create a workflow to extract messages from Twitter, translate and mine them for epidemiological data.

To create a workflow the user starts by adding a Query Selection Service to the Interface Work Area (Fig. 6.6 a). To retrieve messages from Twitter the user connects the Query Selection XML output to the crawler input and specifying the source in the Query Selection input (Fig. 6.6 b). The Query Selection also enables filtering by disease or location so that users can specify specific entities to be searched for.

Users can also text-mine the messages extracted. Since the available text mining service is implemented only for the English language, the user uses a translation service (Fig. 6.6 c). Since the user does not know which language

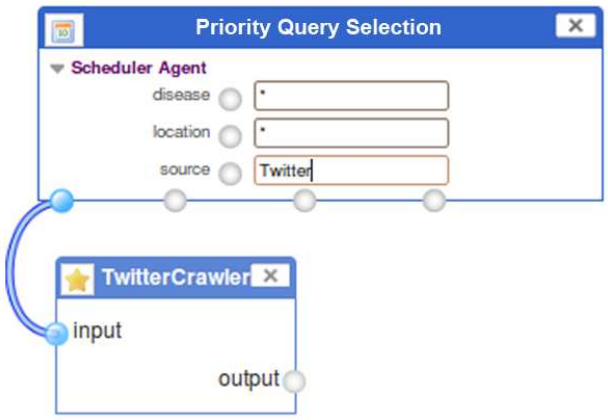


Figure 6.7: Connect the XML output of the Scheduler to the Harvesting Service.

each message is in he/she leaves "input language" blank. The Translation service will identify what language the message is in prior to translating it. The user chooses the desired output language - "en" since he wants the output to be in English - then he/she connects the translation service output to the text mining service (Fig. 6.6 d).

The user can store both the raw messages as well as the occurrences extracted from text mining by connecting both the output of the crawler and the text mining service to a Merge Gate and then connecting it to a Loading Service (Fig. 6.7).

After pressing "Save" on the interface's menu, a JSON message is sent to the BPEL Process Builder, which deploys the process to be run by Apache ODE.

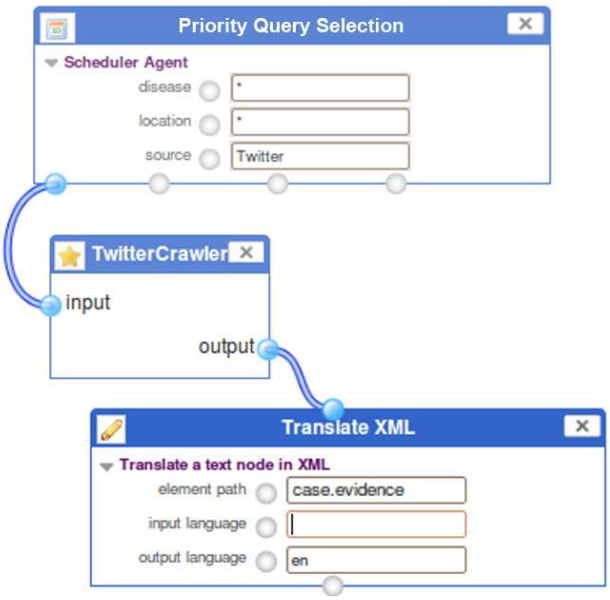


Figure 6.8: To translate the text of the extracted messages use the Translation Service..

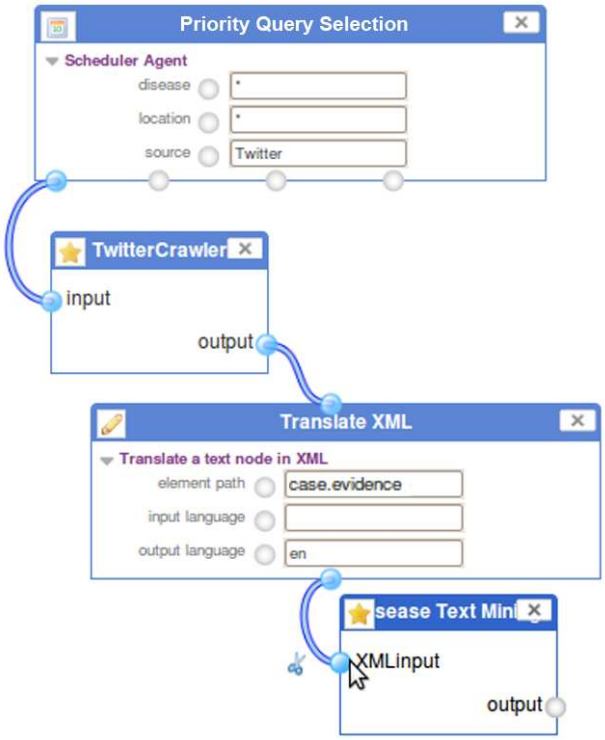


Figure 6.9: After translation a text mining service can be used to extract further information.

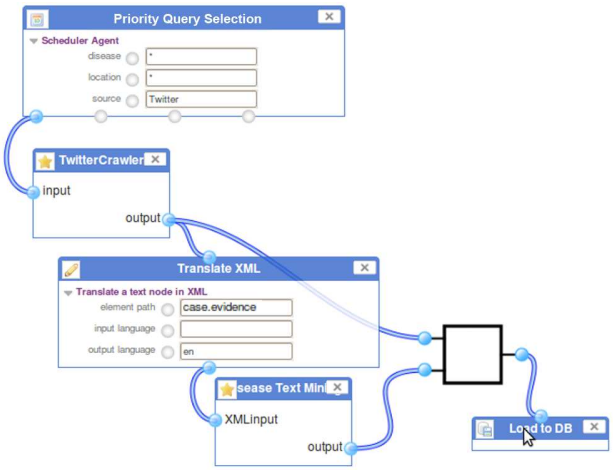


Figure 6.10: Final step of the step-by-step workflow creation. To store both raw messages and text-mined cases connect both outputs to a merge game and then connect it to a loading service.

Chapter 7

Results

In this chapter we summarize the lessons learned with the design and operation of the Epidemic Marketplace, report the statistics gathered (as of July 2013), and provide details on a study that we have conducted with the resources in the platform.

7.1 Lessons Learned

Throughout the four years of the project we have identified a set of requirements for a data repository, and showed how these requirements influenced our decisions for the Epidemic Marketplace. Following these requirements we successfully developed an information platform able to accommodate heterogeneous data resources while fostering data sharing and collaboration.

The Information platform has been improved over the years, since its first release in 2010. By using it in various developments with EPIWORK partners through the monitoring of its use, we have been able to adapt it incrementally to the various needs that have been identified.

Using Fedora Commons as the base repository software proved to be a good initial decission, because it still remains the best all around solution and is fully customizable to suit the needs of our information platform. We have been able to store and manage resources in a large variety of data types and formats. However, browsing and searching the existing resources was still a challenge and some adjustments had to be made to the base version. To facilitate this it was necessary to adapt the repository digital object structure into a "folder-like" structure and enable indexing of resource metadata to make it searcheable.

Accurate metadata is important in an information platform so that users are able to find the content they are looking for. Our use of ontologies in metadata provide users with the ability to describe resources without the ambiguity inherent to free text metadata. Instead of creating a new ontology for the Epidemic Marketplace our solution is NERO which maps terms from a number of ontologies which relate to epidemiology. However, providing non ambiguous metadata is a complex task, and therefore a dedicated user interface provides a set of features to help users to choose the best ontology terms for their resources without having to directly type their URIs.

To foster participation, we additionally included mechanims for user interaction with the repository including a *watch* mechanism which enables users to keep updated on resource changes, the ability to comment on resources to facilitate discussion over data and promote collaboration, and a request mechanism for users who do not find what they are immediately looking for.

The ability to share information in a platform also comes with the responsibility of protecting sensitive data from unauthorized accesses. Access control and data sharing facilities aid in addressing this issue by enabling

users to define permissions over their resources. We use decentralized DAC to give owners the ability to decide who has access to their resources without the interference of administrative staff.

We started from the begining with a distributed architecture, and showeed how to distribute the platform across multiple geographically separated nodes. We later realised that this design is not only important to improve availability and provide scalability, but also the only solution for the integration of heterogenous data from multiple states while addressing all legal and regulatory issues.

The Epidemic Marketplace was designed as information platform that interacts with computational, modeling and forecasting platforms. In order to promote the use of the EM by external applications, we developed a web service API which enables these applications to perform actions over the repository resources. The availability of such API and the simplicity of the interface are key aspects for promoting such integration. And, without such integration, it will be very hard to concentrate a critical mass of updated and relevant datasets.

For human interaction with the platform, we provide a browser-based web interface which enables users to perform all the actions necessary to interact with the EM to create, share, edit, delete, search and request data.

The platform was deployed on the hardware infrastructure through the use of virtualization. Additionally we explored solutions for maximizing the availability of the information platform, which provide a good level of stability to the platform.

7.2 Usage Statistics

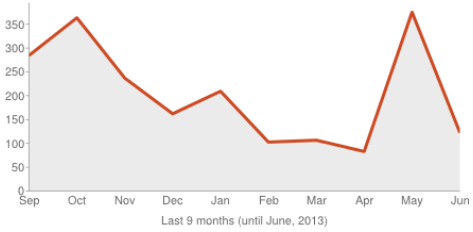
We continued adding resources to the EM repository since it was first released internally, not only to demonstrate the repository functionality and the adopted meta-data schema, but also to demonstrate its integration with the epidemic modeling and visualization software under development in other workpackages of the EPIWORK project. In particular, data harvesting work done in collaboration with the GLEaMviz and Influenzanet teams lead to the creation of new resources.

We present below some of the statistics on the resources and accesses to the Epidemic Marketplace (measured on July 23rd, 2013):

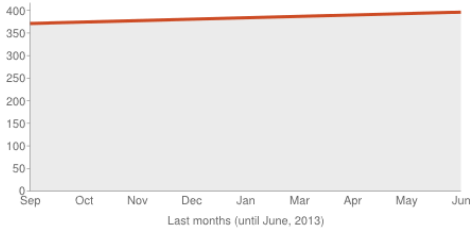
- Number of registered users: 567 (of these 501 result from the integration with GLEaMviz which has seen a boom in activity in the past month)
- Number of distinct uploaders: 27
- Number of pending Resource Requests: 7
- Number of Request views in the EM: 153
- Number of distinct Resources in the EM: 481
- Number of Resources views in the EM: 2577
- Number of visits to the EM Website: 2170 Visits (since July 2011)

To keep track of usage statistics a module was developed for the interface which keeps track of visits, uploaded resources, resource view and request views.

Figure 7.1 contains a few of the usage statistics in last 9 months (until June 2013, as July numbers are not yet complete). These are also available in the website.

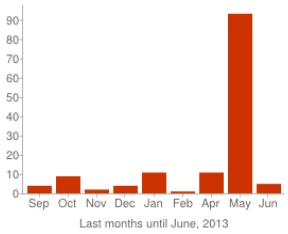


(a) Number of visits for the last 9 months.



(b) Cumulative number of resources for the last 9 months.

(c) Number of resource views for the last 9 months.



(d) Number of request views for the last 9 months.

Figure 7.1: Some usage statistics until June 2013.

7.3 Assessment of Resource Annotations

In the last period of EPIWORK, we conducted a study on a corpus of over a hundred epidemiology scientific articles catalogued in the Epidemic Marketplace. Our goal was elucidating how epidemiological datasets mentioned in the corpus can be properly shared with the community. We investigated the availability of datasets by careful examining the full-text of articles collected from core epidemiology journals for mentions to datasets. For each referred dataset, we assessed the feasibility of semantically characterizing it through annotation with terms from NERO.

We analysed articles published in two top journals in Epidemiology, the Journal of Epidemiology and Community Health (JECH) and Eurosurveillance. For JECH we focused on the period between 1985 and 2011, and selected 34 articles relevant for our purposes. In Eurosurveillance, we read the articles published in the period between October 13, 2011 and August 16, 2012. We found that most of the articles published under Rapid communications, Research articles and Surveillance and outbreak reports contained references to epidemiology datasets, reaching a total of 78 publications.

In the first step of our study one researcher carefully read the full-text articles in this refined set and extracted information about the datasets pertaining to the areas covered in the EM metadata model. Then, the same researcher created, for each article, two types of resources in the EM, one of the Document type, and additional resources of the type Epidemiology dataset to describe the datasets mentioned in the article. All these resources were annotated with the elements of the EM metadata model using the forms provided by the EM online interface. Whenever the exact term was not available in NERO, a broader synonym was selected. The final step consisted

in a second researcher verifying and correcting all the annotations.

In the 112 articles that we read, we identified 112 datasets, one per article, which would be expected given the current standard approach for focusing on self-collected data. Perhaps the most relevant result of our analysis is that none of the articles linked to a data repository where the data could be obtained, or indicated any other means to retrieve it. Any effort to reproduce the research would require contacting the authors to obtain the datasets. However, to truly become useful, these datasets would need to be not only available but also effectively searchable and comparable.

To test the feasibility of semantically describing epidemic resources, we analysed each epidemiology dataset resource in terms of its semantic annotation completion, i.e. we checked if for each EM metadata element, such as disease or drug, there was at least one assigned term, like 'tuberculosis' or 'penincilin'. Figure 7.2 depicts the number of resources that have at least one term for the shown metadata element. All 112 resources were assigned to a disease term, several of which to multiple diseases. Likewise, all were annotated for host. Most resources had information on the used diagnostic method, pathogen involved, and demography. Over half of them mentioned symptoms and transmission mode. However, fewer mentioned drugs, vaccines, vectors or environmental and socioeconomic characteristics. This reveals that the textual descriptions of the datasets are quite complete, since metadata fields that would be considered mandatory (i.e., all epidemiology studies focus on at least one disease affecting at least one kind of host) could all be filled-in, whereas the ones that were less used pertain to more specific characteristics (i.e., not all epidemiology studies focus on vaccination or drug-based treatments).

If on one hand, not all documents provided information to annotate every

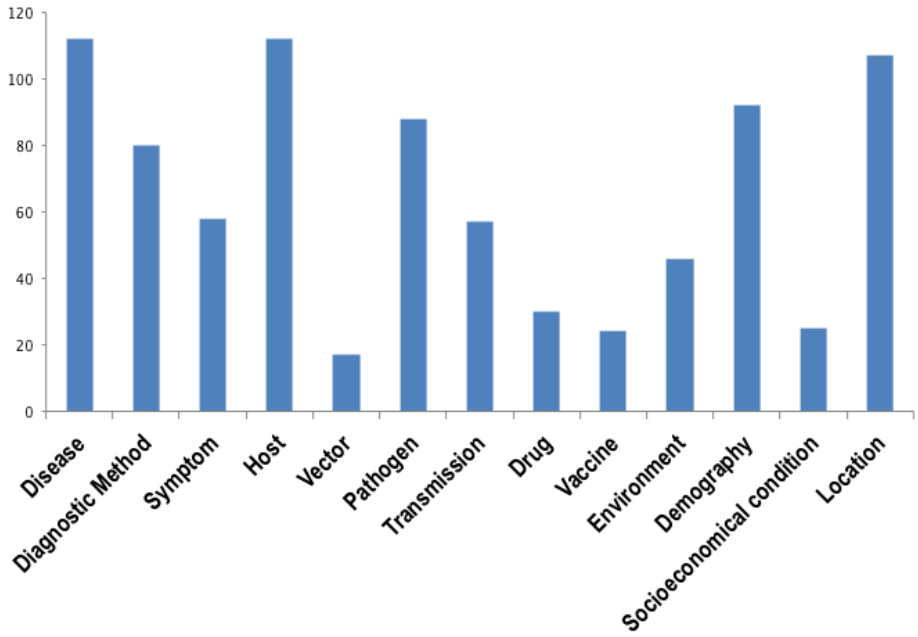


Figure 7.2: Number of resources annotated on each of the EM metadata elements

associated metadata element, in a handful of cases there were no available terms in the vocabularies to adequately describe the datasets in terms of environmental, demographic or socioeconomic conditions. But, by and large, all datasets were accurately characterized with the EM metadata elements, supporting its feasibility for the purpose of semantically annotating epidemic resources.

Ontologies and vocabularies were used to fill-in metadata elements related to the biological information, demography, environment and socioeconomic condition. Figure 7.3 illustrates the distribution of ontologies and vocabularies by each metadata element. NCI Thesaurus and MeSH are used in almost all elements, with the exception of environment for the former, transmission for the latter and vaccines for both. As expected, each one of the ontologies provides the larger coverage for its domain, Disease Ontology for disease, Symptom Ontology for symptom, Transmission Ontology for transmission and Vaccine Ontology for vaccine and, to a lesser extent, the Environment Ontology (ENVO) for environment terms.

Table 7.1 provides an overview of the usage of each NERO resource, both in terms of the total number of annotations and number of unique terms that were used. The most commonly used resource is the NCI Thesaurus, with a total of 469 counts, of which 112 correspond to the assignment of H. sapiens as the host species, followed by the MeSH vocabulary with a total of 215. These correspond to 157 and 132 unique terms, respectively. The Symptom Ontology also has a high number of uses, 188. However, only 58 resources mention symptoms, which is due to the annotation of the same EM metadata element with multiple symptoms. Allowing multiple annotations under the same metadata element is a crucial feature of the EM, since many resources mention multiple diseases, symptoms, drugs, etc. Another interesting fact is

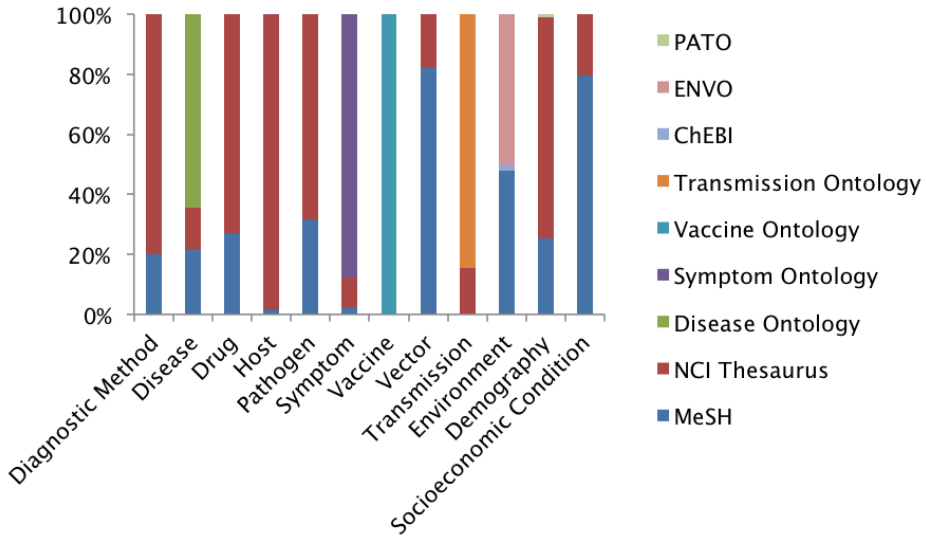


Figure 7.3: Distribution of the terms from each ontology used to fill-in the EM metadata elements

Table 7.1: Number of total annotations and unique terms used from each ontology to annotate an EM dataset resource

Ontologies/vocabularies	**Annotations**	**Unique Terms**
MeSH	215	132
NCI Thesaurus	469	157
Disease Ontology	126	70
Symptom Ontology	188	79
Vaccine Ontology	27	16
Transmission Ontology	49	9
ChEBI	1	T1
ENVO	24	9
PATO	1	1

that many of these resources share the same disease, as shown in Figure 7.4. This is an excellent illustration of the opportunity in crossing information from different datasets belonging to the same disease to support broader studies. Likewise, many of the described resources share symptoms and drugs.

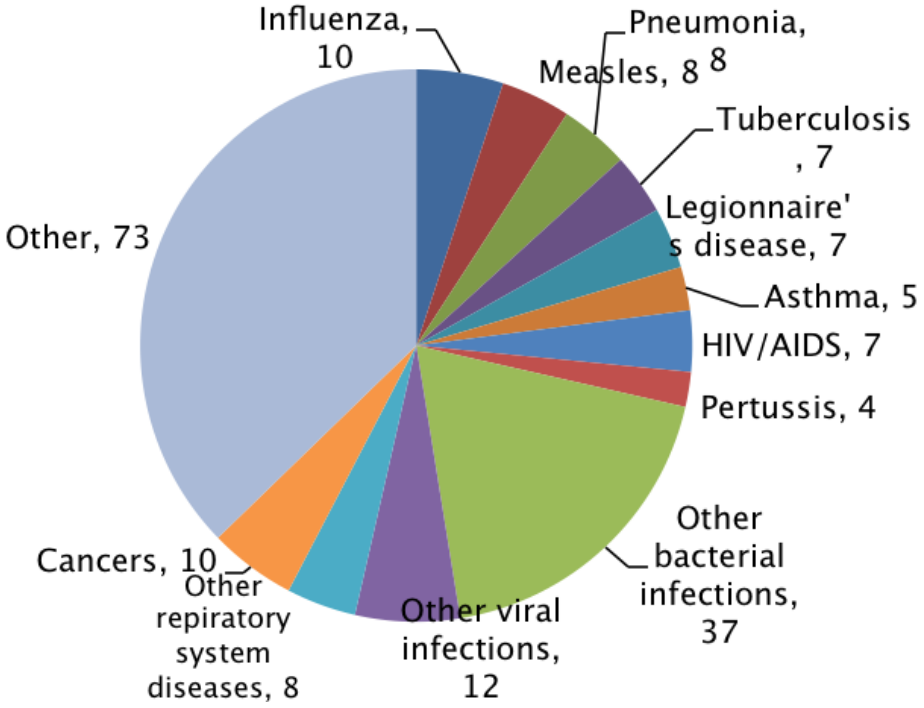


Figure 7.4: Number of EM dataset resources referencing different diseases

The annotation process was performed using the Epidemic Marketplace online interface to fill-in the metadata elements using the information available in the articles that mentioned their use. The analysis demonstrated the feasibility of providing accurate and detailed semantic descriptions for epidemiological datasets. Most of the analysed articles were published in little less than one year in a top epidemiology journal, which is indicative of the high importance of available datasets in epidemiological studies.

Moreover, it also makes the case for adopting a common platform for sharing and describing epidemiology resources, such the EM. By having their descriptions accessible in a centralized manner, epidemiology researchers can easily cross information and retrieve relevant data, which under the current model of manually searching publications often becomes a time consuming and frustrating task.

Chapter 8

Conclusions

In this report we presented the final specification of the Epidemic Marketplace, resulting from the work performed in the 54 month (48 months + 6 months extension) of the EPIWORK project. We have fulfilled the goal of developing an information platform for epidemiological data sharing which can be used interchangeably by users and external applications, including those created by EPIWORK partners.

Although our design is focused on the scope of an epidemiological information platform, the heterogeneous nature of epidemiological content makes us think that the solution presented in this report could be offered as a package which, given some customization, can be applied to other similar problems where finding and sharing resources is the center of the problem.

Overall, the experience of designing and operating the EM for several years, with the collected resources that it now includes, motivates the continuation of efforts towards the adoption of a common platform for sharing and describing epidemiology resources. By having their descriptions accessible in a centralized manner, epidemiology researchers can easily cross information and retrieve relevant data, which under the current model

of manually searching publications often becomes a time consuming and frustrating task.

Such an endeavour is, however, largely dependent on the adoption of these practices by the community. And changing such practices will take much longer than the duration of Epiwork and will demand much more effort. To have an impact in changing the way epidemiology studies are conducted, both the semantic annotation and the sharing of epidemic resources needs to be encouraged. Although good practices in sharing data, actually even just metadata, of epidemiologically relevant datasets would benefit immensely the advancement of the field, it is clear that the right incentives and policies have to be set in order to initially ignite this process. Among those, one of the most easily achievable instruments is to make mandatory the metadata sharing practice for funded projects by national and international funding agencies. In addition, some journals are already requiring authors to make their data available (for instance PLOS journals). However, most authors fail to comply with this requirement, indicating that this strategy is not sufficient [30]. In molecular biology this incentive also comes from the journals themselves, but, instead of data being supplied by authors on demand, it must be submitted to a public database before publication. This has proven to be a more successful strategy and was in great part also fostered by the development of the appropriate infrastructures to manage genomic data online. This somehow also minimizes the common issue in how to give credit to the data authors, since the credit is implicitly derived from the article publication [31].

Due to the sensitive nature of epidemiological data, this strategy cannot be straightforwardly applied. Recent research has confirmed that authors are more reluctant to share data obtained from human subjects studies [32]. However, data access can be managed, as it is in the EM, in a fashion

where although the metadata is in principle accessible by all users, access to the data itself can be protected and restricted to authorized users. This ensures that the data can remain private, while some of the knowledge about the dataset can still be shared. In particular, individual patient data meta-analyses, crucial in 21st century epidemiology, can be greatly aided by the adoption of a data-sharing platform. In a recent study, it was found that most meta-analysts do not even attempt to collect individual patient data either due to a lack of time, resources or believing that the undertaking would be too difficult [33]. Thus, we conclude that of the multiple complex issues preventing the sharing of epidemic resources in integrated platforms, as it has become the norm in other domains, the need for protecting the information that would be required for cataloguing such resources is a myth, as it is already available in the scientific articles based on such resources.

The adoption of an epidemic datasharing platform, such as the Epidemic Marketplace, would greatly support these studies, as well as more traditional research, by contributing to the reproducibility of research in the field and augmenting the knowledge of the researchers therein. Moreover, they would be expedited by the common annotation of these resources under the same metadata model and ontologies, which would greatly simplify their comparison. By having resources annotated with ontologies, we can compute similarities between resources based, for instance, on geographical location or list of symptoms, supporting near instantaneous retrieval of resources ranked by similarity. This would contribute to new avenues of research based on the analysis of similar resources, be it in terms of diseases, population or geographical location.

We believe that we are now beyond the question of "if" or even "when" to share epidemiological data [1], and should be now addressing the "how".

Efforts such as the Epidemic Marketplace and the terminology services by the ECDC [34] present appropriate proposals for addressing this complex but unavoidable issue.

Acknowledgements

The development of the Epidemic Marketplace is a collective work. We have received great feedback in the EPIWORK workshops and project meetings in the past four years. We have been able to integrate Influenzanet and the GLEAMviz platform with the Epidemic Marketplace though a collaborative process with the partners participating in WP4 and WP5. Over the years, several people joined and left the project. The key contributors to the EM development have not authors of this report. The main development team includes, in no particular order: Mário J. Silva, Dulce Domingos, Francisco Couto, Fabrício Silva, Cátia Pesquita, Luís F. Lopes, João Zamite, João Ferreira, Paulo Graça, Hugo Ferreira, Hugo Santos, Vera Carvalho, Tiago Posse, Carlos Sousa, Patrícia Sousa, Juliana Duque Corrado Gioannini and Daniela Paolotti.

Bibliography

[1] T. Hey, S. Tansley, and K. Tolle. (2009) The fourth paradigm: Data-intensive scientific discovery. redmond, wa: Microsoft. [Online] Available: http://research.microsoft.com/en-us/collaboration/fourthparadigm/. [Accessed July, 2013].

[2] J. Wood, T. Anderson, A. Bachem, C. Best, F. Genova, D. Lopez, W. Los, M. Marinucci, L. Romary, H. Van de Sompel *et al.*, "Riding the Wave–How Europe can gain from the rising tide of scientific data," *European Union*, 2010.

[3] S. M., B. L., B. T. J., B. D. D., B. J. S., B. C. *et al.*, "Digital epidemiology," *PLOS Computational Biology*, vol. 8, no. 7, 2012.

[4] W. V. Broeck, C. Gioannini, B. Gonçalves, M. Quaggiotto, V. Colizza, and A. Vespignani, "The gleamviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale," *BMC infectious diseases*, vol. 11, no. 1, p. 37, 2011.

[5] D. L. Chao, M. E. Halloran, V. J. Obenchain, and I. M. Longini, "Flute, a publicly available stochastic influenza epidemic simulation model," *PLoS Computational Biology*, vol. 6, no. 1, 2010.

[6] L. F. Lopes, F. A. Silva, F. Couto, J. Zamite, H. Ferreira, C. Sousa, and M. J. Silva, "Epidemic marketplace: An information management system for epidemiological data." in *Proceedings of the ITBAM - DEXA 2010*, 2010.

[7] Fedora Commons *et al.* (n.d.) Fedora Commons Repository Software Documentation. [Online] Available: https://wiki.duraspace.org/display/FEDORA35/Fedora+3.5+Documentation. [Accessed July, 2013].

[8] W. Yeong, T. Howes, and S. Kille, "Lightweight Directory Access Protocol," 1995.

[9] A. Byron, H. Berry, N. Haug, J. Eaton, J. Walker, and J. Robbins, *Using Drupal*. O'Reilly Media, Incorporated, 2008.

[10] D. C. M. Initiative *et al.* (1999) Dublin core metadata element set, version 1.1: Reference description. [Online] Available: http://dublincore. org/documents/2003/08/26/usageguide. [Accessed July, 2013].

[11] J. D. Ferreira, C. Pesquita, F. M. Couto, and M. J. Silva, "Bringing epidemiology into the semantic web." in *ICBO*, 2012.

[12] T. Berners-Lee and J. Hendler, "Publishing on the semantic web," *Nature*, vol. 410, no. 6832, pp. 1023–1024, 2001.

[13] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 5, no. 3, pp. 1–22, 2009.

[14] J. Zamite, D. Domingos, M. J. Silva, and C. Santos, "Group-based discretionary access control for epidemiological resources," in *HCist 2013 - International Conference on Health and Social Care Information Systems and Technologies*, ser. Procedia Technology. Elsevier, October 2013.

[15] (2013, January) Eprints - digital repository software. [Online]. Available: http://www.eprints.org/

[16] M. Smith, M. Barton, M. Bass, M. Branschofsky, G. McClellan, D. Stuve, R. Tansley, and J. H. Walker, "Dspace: An open source dynamic digital repository," *D-lib Magazine*, vol. 9, no. 1, 2003.

[17] C. Nguyen, J. Dalziel, and S. Cassidy, "Flexible Access Control, Federated Identity and Heterogeneous Metadata Supports for Repositories," *Proceedings of eResearch Australasia 2008*, 2008.

[18] D. Tcsec, "Trusted computer system evaluation criteria," Technical Report 5200.28-STD, US Department of Defense, Tech. Rep., 1985.

[19] R. Thomas, R. Sandhu *et al.*, "Discretionary access control in object-oriented databases: Issues and research directions," in *Proc. 16th National Computer Security Conference*, 1993, pp. 63–74.

[20] A. Grunbacher and A. Nuremberg, "POSIX access control lists on linux," in *Proceedings of the USENIX 2003 Annual Technical Conference, FREENIX track*, 2003, pp. 259–272.

[21] A. Kapica. (2012) Mediawiki Extension:Access Control. [Online] Available: http://www.mediawiki.org/wiki/ Extension:AccessControl. [Accessed July, 2013].

[22] (2013, January) Fcrepo, a client for the fedora commons repository. [Online]. Available: http://pypi.python.org/pypi/fcrepo

[23] N. Shedroff, "Information interaction design: A unified field theory of design," *Information design*, pp. 267–292, 1999.

[24] N. Collier, S. Doan, A. Kawazoe, R. M. Goodwin, M. Conway, Y. Tateno, Q.-H. Ngo, D. Dien, A. Kawtrakul, K. Takeuchi *et al.*, "Biocaster: detecting public health rumors with a web-based text mining system," *Bioinformatics*, vol. 24, no. 24, pp. 2940–2941, 2008.

[25] M. J. Khoury, J. S. Dorman *et al.*, "The human genome epidemiology network," *American journal of epidemiology*, vol. 148, no. 1, pp. 1–3, 1998.

[26] M. Porta, *A dictionary of epidemiology*. Oxford University Press, 2008.

[27] D. A. Lindberg, B. L. Humphreys, and A. T. McCray, "The unified medical language system." *Methods of information in medicine*, vol. 32, no. 4, pp. 281–291, 1993.

[28] C. E. Lipscomb, "Medical subject headings (mesh)," *Bulletin of the Medical Library Association*, vol. 88, no. 3, p. 265, 2000.

[29] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall *et al.*, "The obo foundry: coordinated evolution of ontologies to support biomedical data integration," *Nature biotechnology*, vol. 25, no. 11, pp. 1251–1255, 2007.

[30] C. Savage and A. Vickers, "Empirical study of data sharing by authors publishing," *PloS ONE*, vol. 4, no. 9, 2009.

[31] D. Gardner, A. Toga, G. Ascoli, J. Beatty, J. Brinkley, A. Dale *et al.*, "Towards effective and rewarding data sharing," *Neuroinformatics*, vol. 1, no. 3, pp. 289–295, 2003.

[32] H. Piwowar, "Who shares? who doesn't? factors associated with openly archiving raw research data," *PLoS ONE*, vol. 6, no. 7, 2011.

[33] S. Kovalchik, "Survey finds that most meta-analysts do not attempt to collect individual patient data," *Journal of Clinical Epidemiology*, 2012.

[34] L. Balkányi, G. Héja, and C. Perucha, "Building and using terminology services for the european centre for disease prevention and control," in *Electronic Healthcare*. Springer, 2010, pp. 116–123.